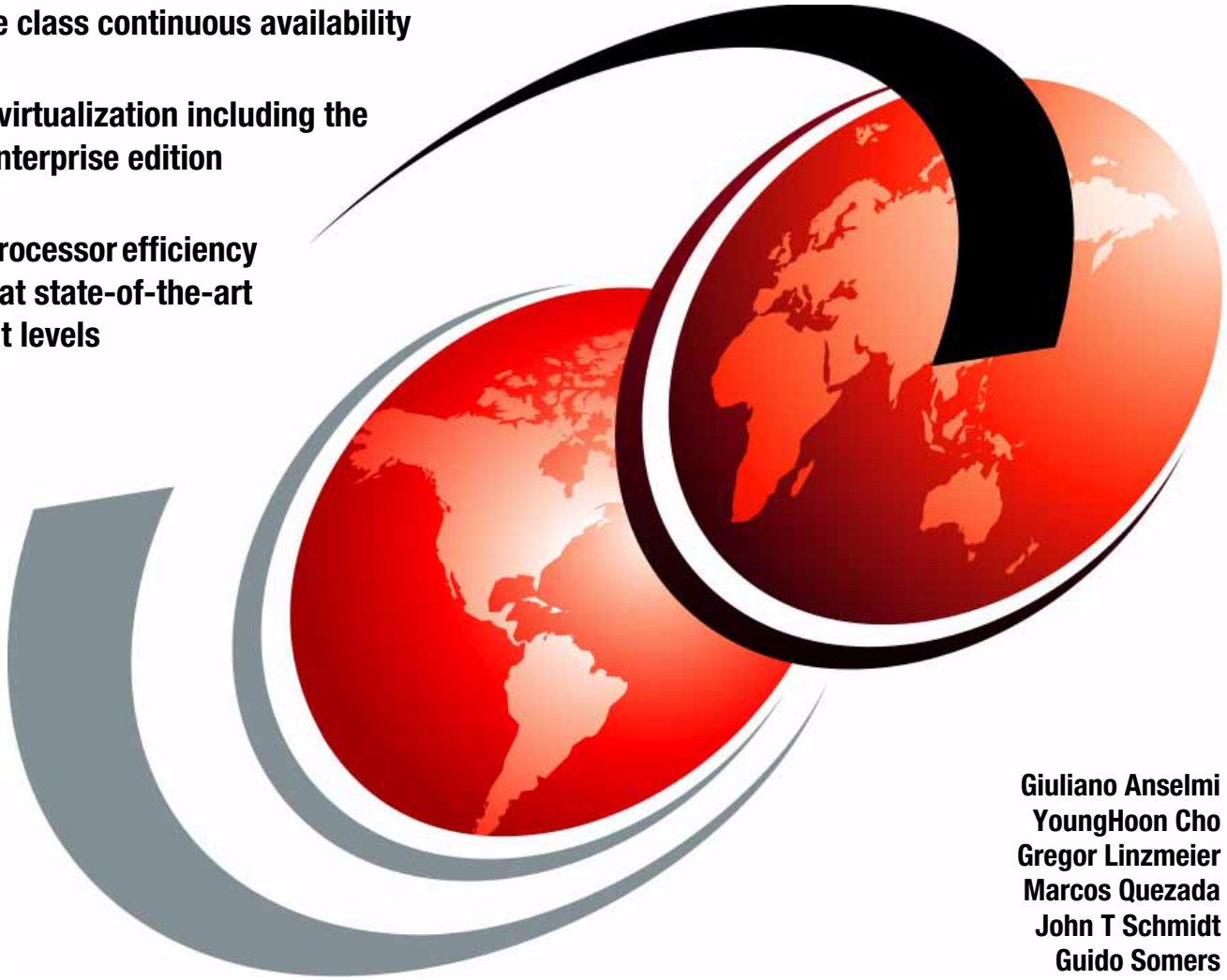
Draft Document for Review May 28, 2009 2:58 pm

REDP-4405-00

IBM Power 570 Technical Overview and Introduction

- Expandable modular design supporting advanced mainframe class continuous availability
- PowerVM virtualization including the optional Enterprise edition
- POWER6 processor efficiency operating at state-of-the-art throughput levels



Giuliano Anselmi
YoungHoon Cho
Gregor Linzmeier
Marcos Quezada
John T Schmidt
Guido Somers

Redpaper

ibm.com/redbooks

International Technical Support Organization

IBM Power 570 Technical Overview and Introduction

October 2008

Note: Before using this information and the product it supports, read the information in “Notices” on page vii.

First Edition (October 2008)

This edition applies to the IBM Power 570 (9117-MMA).

© Copyright International Business Machines Corporation 2008. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices vii
Trademarks viii

Preface ix
The team that wrote this paper ix
Become a published author x
Comments welcome xi

Chapter 1. General description 1
1.1 System specifications 3
1.2 Physical package 3
1.3 System features 4
1.3.1 Processor card features 5
1.3.2 Memory features 6
1.3.3 Disk and media features 7
1.3.4 I/O drawers 8
1.3.5 Hardware Management Console models 12
1.4 System racks 12
1.4.1 IBM 7014 Model T00 rack 13
1.4.2 IBM 7014 Model T42 rack 14
1.4.3 The AC power distribution unit and rack content 14
1.4.4 Intelligent Power Distribution Unit (iPDU) 16
1.4.5 Rack-mounting rules 17
1.4.6 Useful rack additions 18
1.4.7 OEM rack 19

Chapter 2. Architecture and technical overview 23
2.1 The POWER6 processor 24
2.1.1 Decimal floating point 26
2.1.2 AltiVec and Single Instruction, Multiple Data 27
2.2 IBM EnergyScale technology 27
2.2.1 Hardware and software requirements 28
2.3 Processor cards 29
2.3.1 Processor drawer interconnect cables 30
2.3.2 Processor clock rate 30
2.4 Memory subsystem 31
2.4.1 Fully buffered DIMM 31
2.4.2 Memory placements rules 32
2.4.3 Memory consideration for model migration from p5 570 to 570 33
2.4.4 OEM memory 34
2.4.5 Memory throughput 35
2.5 System buses 35
2.5.1 I/O buses and GX+ card 35
2.5.2 Service processor bus 36
2.6 Internal I/O subsystem 37
2.6.1 System ports 38
2.7 Integrated Virtual Ethernet adapter 38
2.7.1 Physical ports and system integration 39
2.7.2 Feature code port and cable support 41

© Copyright IBM Corp. 2008. All rights reserved.

2.7.3 IVE subsystem 42
2.8 PCI adapters 43
2.8.1 LAN adapters 44
2.8.2 SCSI and SAS adapters 44
2.8.3 iSCSI 45
2.8.4 Fibre Channel adapter 47
2.8.5 Graphic accelerators 47
2.8.6 Asynchronous PCI adapters 47
2.8.7 Additional support for existing PCI adapters 48
2.9 Internal storage 48
2.9.1 Integrated RAID options 48
2.9.2 Split backplane 49
2.9.3 Internal media devices 49
2.9.4 Internal hot-swappable SAS drives 49
2.10 External I/O subsystems 50
2.10.1 7311 Model D11 I/O drawers 50
2.10.2 Consideration for 7311 Model D10 I/O drawer 51
2.10.3 7311 Model D20 I/O drawer 52
2.10.4 7311 Model D11 and Model D20 I/O drawers and RIO-2 cabling 53
2.10.5 7311 I/O drawer and SPCN cabling 54
2.10.6 7314 Model G30 I/O drawer 54
2.11 External disk subsystems 55
2.11.1 IBM System Storage EXP 12S (FC 5886) 56
2.11.2 IBM TotalStorage EXP24 Expandable Storage 57
2.11.3 IBM System Storage N3000, N5000 and N7000 58
2.11.4 IBM TotalStorage Storage DS4000 Series 58
2.11.5 IBM TotalStorage Enterprise Storage Server 58
2.12 Hardware Management Console 59
2.12.1 High availability using the HMC 61
2.12.2 Operating System Support 62
2.13 Service information 65
2.13.1 Touch point colors 65
2.13.2 Operator Panel 65
2.14 System firmware 67
2.14.1 Service processor 70
2.14.2 Redundant service processor 71
2.14.3 Hardware management user interfaces 71

Chapter 3. Virtualization 75
3.1 POWER Hypervisor 76
3.2 Logical partitioning 78
3.2.1 Dynamic logical partitioning 78
3.2.2 Micro-Partitioning 78
3.2.3 Processing mode 79
3.3 PowerVM 81
3.3.1 PowerVM editions 81
3.3.2 Virtual I/O Server 81
3.3.3 PowerVM Lx86 85
3.3.4 PowerVM Live Partition Mobility 87
3.3.5 PowerVM AIX 6 Workload Partitions 88
3.3.6 PowerVM AIX 6 Workload Partition Manager 89
3.3.7 Operating System support for PowerVM 90

3.4 System Planning Tool 91

Chapter 4. Continuous availability and manageability 93
4.1 Reliability 93
4.1.1 Designed for reliability 93
4.1.2 Placement of components 94
4.1.3 Redundant components and concurrent repair 94
4.1.4 Continuous field monitoring 95
4.2 Availability 95
4.2.1 Detecting and deallocating failing components 95
4.2.2 Special uncorrectable error handling 101
4.2.3 Cache protection mechanisms 102
4.2.4 PCI Error Recovery 103
4.3 Serviceability 104
4.3.1 Detecting errors 105
4.3.2 Diagnosing problems 107
4.3.3 Reporting problems 109
4.3.4 Notifying the appropriate contacts 110
4.3.5 Locating and repairing the problem 111
4.4 Operating System support for RAS features 113
4.5 Manageability 116
4.5.1 Service processor 116
4.5.2 System diagnostics 116
4.5.3 Electronic Service Agent 117
4.5.4 Manage serviceable events with the HMC 119
4.5.5 Hardware user interfaces 119
4.5.6 IBM System p firmware maintenance 120
4.5.7 Management Edition for AIX 121
4.5.8 IBM Director 122
4.6 Cluster solution 122

Related publications 125
IBM Redbooks 125
Other publications 125
Online resources 126
How to get Redbooks 126
Help from IBM 126

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

© Copyright IBM Corp. 2008. All rights reserved.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Redbooks (logo) ®
Eserver®
eServer™
iSeries®
i5/OS®
pSeries®
AIX 5L™
AIX®
Chipkill™
DS4000™
DS6000™
DS8000™
Electronic Service Agent™
EnergyScale™
Enterprise Storage Server®
HACMP™
IntelliStation®
IBM Systems Director Active Energy Manager™
IBM®
Micro-Partitioning™
OpenPower®
Power Architecture®
PowerPC®
PowerVM™
Predictive Failure Analysis®
POWER™
POWER Hypervisor™
POWER4™
POWER5™
POWER5+™
POWER6™
Redbooks®
RS/6000®
System i™
System i5™
System p™
System p5™
System x™
System z™
System Storage™
Tivoli®
TotalStorage®
Workload Partitions Manager™
1350™

The following terms are trademarks of other companies:

ABAP, SAP NetWeaver, SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

Oracle, JD Edwards, PeopleSoft, Siebel, and TopLink are registered trademarks of Oracle Corporation and/or its affiliates.

InfiniBand, and the InfiniBand design marks are trademarks and/or service marks of the InfiniBand Trade Association.

Flex, and Portable Document Format (PDF) are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, other countries, or both.

Java, and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Internet Explorer, Microsoft, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM® Redpaper is a comprehensive guide covering the IBM Power™ 570 server supporting AIX, IBM i, and Linux for Power operating systems. The goal of this paper is to introduce the major innovative Power 570 offerings and their prominent functions, including the following.

- Unique modular server packaging
- New POWER6™ processors available at frequencies of 4.2, 4.4, and 5.0 GHz.
- The POWER6 processor available at frequencies of 3.5, 4.2, and 4.7 GHz.
- The specialized POWER6 DDR2 memory that provides greater bandwidth, capacity, and reliability.
- The 1 Gb or 10 Gb Integrated Virtual Ethernet adapter that brings native hardware virtualization to this server
- PowerVM™ virtualization including PowerVM Live Partition Mobility
- Redundant service processors to achieve continuous availability

Professionals wishing to acquire a better understanding of IBM System p products should read this Redpaper. The intended audience includes:

- Clients
- Sales and marketing professionals
- Technical support professionals
- IBM Business Partners
- Independent software vendors

This Redpaper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the 570 system.

This Redpaper does not replace the latest marketing materials and tools. It is intended as an additional source of information that, together with existing materials, may be used to enhance your knowledge of IBM server solutions.

The team that wrote this paper

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, Austin Center.

Giuliano Anselmi works with passion for IBM devoted to RS/6000® and pSeries® systems since 15 years, having a very deep knowledge of the related hardware, and solutions. He used to be a pSeries Systems Product Engineer for 7 years, supporting Web Server Sales Organization, IBM Sales, Business Partners, Technical Support Organizations. In 2004, he joined Field Technical Sales Support group and has been accredited as IT specialist in 2007. Currently he plays the role of system architect in IBM STG and supporting the General Business division.

YoungHoon Cho is a System p Product Engineer at the pSeries post-sales Technical Support Team in IBM Korea. He has seven years of experience working on RS/6000 and

© Copyright IBM Corp. 2008. All rights reserved.

System p products. He is an IBM Certified Specialist in System p and AIX® 5L™. He provides second line support to field engineers with technical support on System p, and system management.

Gregor Linzmeier is an IBM Advisory IT Specialist for IBM System p workstation and entry servers as part of the Systems and Technology Group in Mainz, Germany supporting IBM sales, Business Partners, and clients with pre-sales consultation and implementation of client/server environments. He has worked for more than 15 years as an infrastructure specialist for RT, RS/6000, IBM IntelliStation® POWER™, and AIX in large CATIA client/server projects. Actual engagements are AIX Thin Server, Partition migration and Green IT.

Marcos Quezada is a Senior Accredited IT Specialist in Argentina. He has 10 years of IT experience as a UNIX systems' pre-sales specialist and as a Web Project Manager. He holds a degree on Informatics Engineering from Fundación Universidad de Belgrano. His areas of expertise include IBM RS/6000, IBM eServer™ pSeries/p5 and Power Systems servers under the AIX operating system and pre-sales support of IBM Software, SAP® and Oracle® solutions architecture running on IBM UNIX Systems with focus on competitive accounts..

John T Schmidt is an Accredited IT Specialist for IBM and has over 7 years experience with IBM and System p. He has a degree in Electrical Engineering from the University of Missouri - Rolla and an MBA from Washington University in St. Louis. He is currently working in the United States as a presales Field Technical Sales Specialist for System p in St. Louis, MO.

Guido Somers is a Cross Systems Certified IT Specialist working for IBM Belgium. He has 13 years of experience in the Information Technology field, ten years of which were within IBM. He holds degrees in Biotechnology, Business Administration, Chemistry, and Electronics, and did research in the field of Theoretical Physics. His areas of expertise include AIX, Linux®, system performance and tuning, logical partitioning, virtualization, HACMP™, SAN, IBM System p servers, as well as other IBM hardware offerings. He currently works as a Client IT Architect for Infrastructure and Global ISV Solutions in the e-Business Solutions Technical Support (eTS) organization. He is also the author of the second edition of *Integrated Virtualization Manager on IBM System p5*™, REDP-4061, and the *PowerVM Live Partition Mobility on IBM System p*, SG24-740.

The project that produced this publication was managed by:
Scott Vetter, PMP

Thanks to the following people for their contributions to this project:

George Ahrens, Ron Arroyo, Brad Behle, Nick Bofferding, Martha Broyles, Pat Buckland, Curtis Eide, Chris Eisenmann, Michael S. Floyd, Chris Francois, Andrew J. Geissler, Gordon Grout, Volker Haug, Daniel J. Henderson, Tenley Jackson, Robert G. Kovacs, Hye-Young McCreary, Bill Mihaltse, Jim A. Mitchell, Thoi Nguyen, Amartey Pearson, Cale Rath, Todd Rosedahl, Terry Schardt, Julissa Villarreal, Brian Warner, Christine I. Wang.
IBM US

Bruno Digiovani
IBM Argentina

Become a published author

Join us for a two- to six-week residency program! Help write a book dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You

will have the opportunity to team with IBM technical professionals, Business Partners, and Clients.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you will develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks® in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an e-mail to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

1

General description

The innovative IBM Power 570 mid-range server with POWER6 and available POWER6+ f processor cards delivers outstanding price/performance, mainframe-inspired reliability and availability features, flexible capacity upgrades, and innovative virtualization technologies to enable management of growth, complexity, and risk.

The Power 570 leverages your exisiting investments by supporting AIX, IBM i, and Linux for Power, and x86 Linux applications on a single server. It is available in 2-, 4-, 8-, 12-, and 16-core and 32-core configurations. As with the p5 570, the POWER6-based 570s modular symmetric multiprocessor (SMP) architecture is constructed using 4U (EIA units), 4-core or 8-core building block modules (also referred to as nodes, or CECs). Each of these nodes supports four POWER6 3.5, 4.2 or 4.7 GHz dual-core processors, and new POWER6 4.2 GHz dual-core processors, or POWER6+ 4.4, and 5.0 GHz four-core processors along with cache, memory, media, disks, I/O adapters, and power and cooling to create a balanced, extremely high-performance rack-mount system.

This design allows up to four modules to be configured in a 19-inch rack as a single SMP server, allowing clients to start with what they need and grow by adding additional building blocks. A fully configured 570 server may consist of 32 processor cores, 768 GB of DDR2 memory, four media bays, integrated ports for attaching communications devices, 24 mixed PCI-X and PCI Express adapter slots, and 24 internal SAS (Serial Attached SCSI) drives accommodating up to 7.2 TB of internal disk storage.

The 64-bit POWER6 processors in this server are integrated into a dual-core single chip module and a dual-core dual chip module, with 32 MB of L3 cache, 8 MB of L2 cache, and 12 DDR2 memory DIMM slots. The unique DDR2 memory uses a new memory architecture to provide greater bandwidth and capacity. This enables operating at a higher data rate for large memory configurations. Each new processor card can support up to 12 DDR2 DIMMs running at speeds of up to 667 MHz.

As with the POWER5™ processor, simultaneous miltithreading enabling two threads to be executed at the same time on a single processor core is a standard feature of POWER6 technology. Introduced with the POWER6 processor design is hardware decimal floating-point support improving the performance of the basic mathematical calculations of financial transactions that occur regularly on today's business computers. The POWER6

© Copyright IBM Corp. 2008. All rights reserved.

processor also includes an AltiVec SIMD accelerator, which helps to improve the performance of high performance computing (HPC) workloads.

All Power Systems servers can utilize logical partitioning (LPAR) technology implemented using System p virtualization technologies, the operating system (OS), and a hardware management console (HMC). Dynamic LPAR allows clients to dynamically allocate many system resources to application partitions without rebooting, allowing up to 16 dedicated processor partitions on a fully configured system.

In addition to the base virtualization that is standard on every System p server, two optional virtualization features are available on the server: PowerVM Standard Edition (formerly Advanced POWER Virtualization (APV) Standard) and PowerVM Enterprise Edition (formerly APV Enterprise).

PowerVM Standard Edition includes IBM Micro-Partitioning™ and Virtual I/O Server (VIOS) capabilities. Micro-partitions can be defined as small as 1/10th of a processor and be changed in increments as small as 1/100th of a processor. Up to 160 micro-partitions may be created on a 16-core 570 system. VIOS allows for the sharing of disk and optical devices and communications and Fibre Channel adapters. Also included is support for Multiple Shared Processor Pools and Shared Dedicated Capacity.

PowerVM Enterprise Edition includes all features of PowerVM Standard Edition plus Live Partition Mobility, newly available with POWER6 systems. It is designed to allow a partition to be relocated from one server to another while end users are using applications running in the partition.

Other features introduced with POWER6 processor-based technology include an Integrated Virtual Ethernet adapter standard with every system, the Processor Instruction Retry feature automatically monitoring the POWER6 processor and, if needed, restarting the processor workload without disruption to the application, and a new HMC (Hardware Management Console) graphical user interface offering enhanced systems control.

1.1 System specifications

Table 1-1 lists the general system specifications of a single Central Electronics Complex (CEC) enclosure.

Table 1-1 System specifications

Description	Range (operating)
Operating temperature	5 to 35 degrees C (41 to 95 F)
Relative humidity	8% to 80%
Maximum wet bulb	23 degrees C (73 F)
Noise level	► with 3.5 GHz processors FC 5620: 7.1 bels ► with 3.5 GHz processors FC 5620 and acoustic rack doors: 6.7 bels ► with 4.2 GHz processors FC 5622: 7.1 bels ► with 4.2 GHz processors FC 5622 and acoustic rack doors: 6.7 bels ► with 4.7 GHz processors FC 7380: 7.4 bels ► with 4.7 GHz processors FC 7380 and acoustic rack doors: 6.9 bels
Operating voltage	200 to 240 V ac 50/60 Hz
Maximum power consumption	1400 watts (maximum)
Maximum power source loading	1.428 kVA (maximum)
Maximum thermal output	4778 BTU[a]/hr (maximum)
Maximum altitude	3,048 m (10,000 ft)

a. British Termal Unit (BTU)

1.2 Physical package

The system is available only in a rack-mounted form factor. It is a modular-built system utilizing between one and four building block enclosures. Each of these CEC drawer building blocks is packaged in a 4U[1] rack-mounted enclosure. The major physical attributes for each building block are shown in Table 1-2.

Table 1-2 Physical packaging of CEC drawer

Dimension	One CEC drawer
Height	174 mm (6.85 in)
Width	483 mm (19.0 in.)
Depth	► 824 mm (32.4 in.) from front of Bezel to rear of Power Supply ► 674 mm (25.6 in.) front rack rail mounting surface to I/O Adapter Bulkhead ► 793 mm (31.2 in.) front rack rail mounting surface to rear of Power Supply
Weight	63.6 kg (140 lb.)

To help ensure the installation and serviceability in non-IBM, industry-standard racks, review the vendor's installation planning information for any product-specific installation requirements.

[1] One Electronic Industries Association Unit (1U) is 44.45 mm (1.75 in.).

Figure 1-1 shows system views.

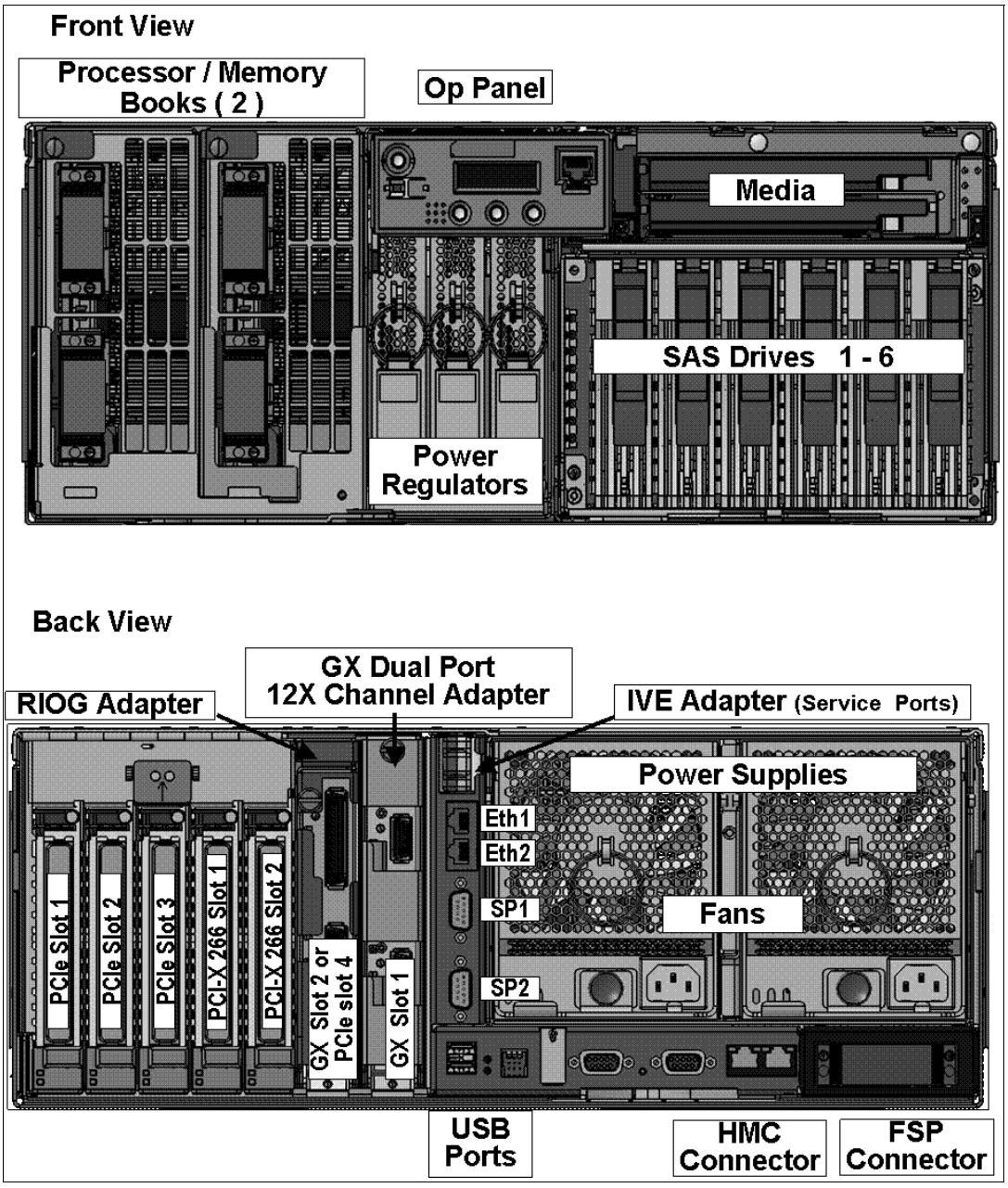


Figure 1-1 Views of the system

1.3 System features

The full system configuration is made of four CEC building blocks. It features:

- 2-, 4-, 8-, 12-, 16-, and 32-core configurations utilizing the POWER6 chip on up to eight dual core processor cards, or eight dual-core POWER6 dual-chip processor cards.
- Up to 192 GB DDR2 memory per enclosure, 768 GB DDR2 max per system. Available memory features are 667 MHz, 533 MHz, or 400 MHz depending on memory density.
- Up to 6 SAS DASD disk drives per enclosure, 24 max per system.

- 6 PCI slots per enclosure: 4 PCIe, 2 PCI-X; 24 PCI per system: 16 PCIe, 8 PCI-X.
- Up to 2 GX+ adapters per enclosure; 8 per system
- One hot-plug slim-line media bay per enclosure, 4 max per system.

The external processor fabric bus in this system is modular. For a multiple-drawer server configuration, a processor fabric cable or cables, and a service interface cable are required. Cable features are available for connecting pairs of drawers, three drawer stacks, and four drawer stacks. With this modular approach, a separate cable is required to connect each drawer to each other drawer in a multi-enclosure stack (See 2.2.1 and 2.4.2).

The service processor (SP), which is described in 2.14.1, “Service processor” on page 70.

Each system includes the following native ports:

- Choice of integrated (IVE) I/O options -- one per enclosure.
  - 2-port 1 Gigabit Integrated Virtual Ethernet controller with two system ports (10/100/1000 twisted pair).
  - 4-port 1 Gigabit Integrated Virtual Ethernet controller with one system port (10/100/1000 twisted pair).
  - 2-port 10 Gigabit Integrated Virtual Ethernet controller (SR optical) with one system port.
- Two USB ports per enclosure.
- Two system (serial) ports per enclosure. Only the ports in the base enclosure are active, and only when an HMC is not attached.
- Two HMC ports per enclosure. The HMC must be attached to CEC enclosure 1 (and CEC enclosure 2 to support redundant Service Processors).
- Two SPCN ports per enclosure.

In addition, each building block features one internal SAS controller, redundant hot-swappable cooling fans, redundant power supplies, and redundant processor voltage regulators.

1.3.1 Processor card features

Each of the four system enclosures has two processor sockets and can contain two POWER6/POWER6+ dual-core 64-bit processor card features, or two POWER6 dual-core dual-chip processor card features. They are configured as dual cores on a single chip module or dual chip module with 32 MB of L3 cache, 8 MB of L2 cache, and 12 DDR2 memory DIMM slots.

The POWER6 processor is available at frequencies of 3.5, 4.2, or 4.7 GHz. The POWER6+ processor is available at frequencies of 4.2, 4.4, and 5.0 GHz.Each system must have a minimum of two active processors. A system with one enclosure may have one or two processor cards installed. A system with two, three, or four enclosures must have two processor cards in each enclosure. When two or more processor cards are installed in a system, all cards must have the same feature number.

All processor card features are available only as Capacity on Demand (CoD). The initial order of the system must contain the feature code (FC) related to the desired processor card, and it must contain the processor activation feature code. The types of CoD supported are:

- *Capacity Upgrade on Demand (CUoD)* allows you to purchase additional permanent processor or memory capacity and dynamically activate them when needed.

- *Utility CoD* autonomically provides additional processor performance on a temporary basis within the shared processor pool in one minute increments. It adds additional cores to allow greater parellel operation, and can increase the effective L2 cache of the shared processor pool
- *On/Off CoD* enables processors or memory to be temporarily activated in full-day increments as needed.
- *Trial CoD* (exception) offers a one-time, no-additional-charge 30-day trial that allows you to explore the uses of all inactive processor capacity on your server.
- *Trial CoD* (standard) offers a one-time 2-core activation for 30 days.
- *Capacity Backup* (IBM i only) offers a 1 license entitlement to a backup system on a temparary basis.

Table 1-3 contains the feature codes for processor cards at the time of writing.

Table 1-3 Processor card and CoD feature codes

Processor card FC	Description
5620 ► 5670 ► 5640 ► 5650	3.5 GHz Proc Card, 0/2 Core POWER6, 12 DDR2 Memory Slots ► One Processor Activation for Processor FC 5620 ► Utility Billing for FC 5620-100 processor minutes ► On/Off Processor Day Billing for FC 5620
5622 ► 5672 ► 5641 ► 5653	4.2 GHz Proc Card, 0/2 Core POWER6, 12 DDR2 Memory Slots ► One Processor Activation for Processor FC 5622 ► Utility Billing for FC 5622-100 processor minutes ► On/Off Processor Day Billing for FC 5621 or FC 5622
7380 ► 5403 ► 5404 ► 5656	4.7 GHz Proc Card, 0/2 Core POWER6, 12 DDR2 Memory Slots ► One Processor Activation for Processor FC 7380 ► Utility Billing for FC 7380-100 processor minutes ► On/Off Processor Day Billing for FC 7380
7951	On/Off Processor Enablement. This feature can be ordered to enable your server for On/Off Capacity on Demand. Once enabled, you can request processors on a temporary basis. You must sign an On/Off Capacity on Demand contract before you order this feature

1.3.2 Memory features

Processor card feature codes 7380, 5620, and 5622 have 12 memory DIMM slots and must be populated with POWER6 DDR2 Memory DIMMs. Each processor card feature must have a minimum of four DIMMs installed. This includes inactive processor card features present in the system. Table 1-4 shows the memory feature codes that are available at the time of writing.

All memory card features are available only as Capacity on Demand and support the same CoD options described for processors (with the exception of Utility CoD).

- *Memory Trial CoD* (exception) offers a one-time, no-additional-charge 30-day trial that allows you to explore the uses of all memory capacity on your server.
- *Memory Trial CoD* (standard) offers a one-time 4 GB activation for 30 days.

All POWER6 memory features must be purchased with sufficient permanent memory activation features so that each memory feature is at least 50% active, except memory feature code 8129 which must be purchased with Activation feature code 5681 for 100% activation.

Table 1-4 Memory feature codes

Feature code	Description
5692	0/2 GB DDR2 Memory (4X0.5 GB) DIMMs-667 MHz-POWER6 Memory
5693	0/4 GB DDR2 Memory (4X1 GB) DIMMs-667 MHz-POWER6 Memory
5694	0/8 GB DDR2 Memory (4X2 GB) DIMMs-667 MHz-POWER6 Memory
5695	0/16 GB DDR2 Memory (4X4 GB) DIMMs-533 MHz-POWER6 Memory
5696	0/32 GB DDR2 Memory (4X8 GB) DIMMs-400 MHz-POWER6 Memory
5680	Activation of 1 GB DDR2 POWER6 Memory
5691	ON/OFF, 1 GB-1 Day, Memory Billing-POWER6 Memory
7954	On/Off Memory Enablement
8129	0/256 GB DDR2 Memory (32X8 GB) DIMMS- 400 MHz- POWER6 Memory
5681	Activation of 256 GB DDR2 POWER6 Memory

Memory feature codes 5692, 5693, 5694, and 5695 can be mixed on the same POWER6 processor card. Memory feature codes 5696 and 8129 may not be mixed with any other memory feature on a single processor card. A processor card with memory feature 5696 or 8129 can be mixed in the same CEC enclosure with a processor card containing other POWER6 memory features. For all processors and all system configurations, if memory features in a single system have different frequencies, all memory in the system will function according to the lowest frequency present. Memory features 5696 and 8129 cannot be used on processor card feature code 5620.

For all processors and all system configurations, if memory features in a single system have different frequencies, all memory in the system will function according to the lowest frequency present

1.3.3 Disk and media features

Each system building block features one SAS DASD controller with six hot-swappable 3.5-inch SAS disk bays and one hot-plug, slim-line media bay per enclosure. Only the new SAS DASD hard disks are supported internally. The older SCSI DASD hard files can be attached, but must be located in a remote I/O drawer. In a full configuration with four connected building blocks, the combined system supports up to 24 disk bays

Table 1-5 shows the disk drive feature codes that each bay can contain.

Table 1-5 Disk drive feature code description

Feature code	Description
3646	73 GB 15 K RPM SAS Disk Drive
3647	146 GB 15 K RPM SAS Disk Drive
3648	300 GB 15 K RPM SAS Disk Drive

In a full configuration with four connected, the combined system supports up to four media devices with Media Enclosure and Backplane feature 5629.

Any combination of the following DVD-ROM and DVD-RAM drives can be installed:

- FC 5756 IDE Slimline DVD-ROM Drive

- FC 5757 IBM 4.7 GB IDE Slimline DVD-RAM Drive

1.3.4 I/O drawers

The system has seven I/O expansion slots per enclosure, including one dedicated GX+ slot. The other 6 slots support PCI adapters. There are 3 PCIe 8X long slots and 1 PCIe 8X short slot. The short PCIe slot may also be used for a second GX+ adapter. The remaining 2 slots PCI-X long slots. If more PCI slots are needed, such as to extend the number of LPARs, up to 20 I/O drawers on a RIO-2 interface (7311-D11 or 7311-D20) ,and up to 32 I/O drawers on a 12X Channel interface (7314-G30) can be attached.

The adapters that are used in the GX expansion slots are concurrently maintainable on systems with firmware level FM320_xxx_xxx, or later. If the GX adapter were to fail, the card could be replaced with a working card without powering down the system.

7311 Model D11 I/O drawer

The 7311 Model D11 I/O drawer features six long PCI-X slots. Blind-swap cassettes are (FC 7862) are utilized. Two 7311 Model D11 I/O drawers fit side-by-side in the 4U enclosure (FC 7311) mounted in a 19-inch rack, such as the IBM 7014-T00 or 7014-T42.

The 7311 Model D11 I/O drawer offers a modular growth path for systems with increasing I/O requirements. A fully configured system supports 20 attached 7311 Model D11 I/O drawers. The combined system supports up to 128 PCI-X adapters and 16 PCIe adapters. In a full configuration, Remote I/O expansion cards (FC 1800 - GX Dual Port RIO-2) are required.

The I/O drawer has the following attributes:

- 4U rack-mount enclosure (FC 7311) that can hold one or two D11 drawers
- Six PCI-X slots: 3.3 V, keyed, 133 MHz blind-swap hot-plug
- Default redundant hot-plug power and cooling devices
- Two RIO-2 and two SPCN ports

7311 Model D11 I/O drawer physical package

Because the 7311 Model D11 I/O drawer must be mounted into the rack enclosure (FC 7311), these are the physical characteristics of one I/O drawer or two I/O drawers side-by-side:

- One 7311 Model D11 I/O drawer
  - Width: 223 mm (8.8 in.)
  - Depth: 711 mm (28.0 in.)
  - Height: 175 mm (6.9 in.)
  - Weight: 19.6 kg (43 lb.)
- Two I/O drawers in a 7311 rack-mounted enclosure have the following characteristics:
  - Width: 445 mm (17.5 in.)
  - Depth: 711 mm (28.0 in.)
  - Height: 175 mm (6.9 in.)
  - Weight: 39.1 kg (86 lb.)

7311 Model D20 I/O drawer

The 7311 Model D20 I/O drawer is a 4U full-size drawer, which must be mounted in a rack. It features seven hot-pluggable PCI-X slots and, optionally up to 12 hot-swappable disks arranged in two 6-packs. Redundant concurrently maintainable power and cooling is an

optional feature (FC 6268). The 7311 Model D20 I/O drawer offers a modular growth path for systems with increasing I/O requirements. When fully configured with 20 attached 7311 Model D20 drawers, the combined system supports up to 148 PCI-X adapters, 16 PCIe adapters, and 264 hot-swappable disks. In a full configuration, Remote I/O expansion cards (FC 1800 - GX Dual Port RIO-2) are required.

PCI-X and PCI cards are inserted into the slots from the top of the I/O drawer. The installed adapters are protected by plastic separators, which are designed to prevent grounding and damage when adding or removing adapters.

The drawer has the following attributes:

- 4U rack mount enclosure assembly
- Seven PCI-X slots: 3.3 V, keyed, 133 MHz hot-plug
- Two 6-pack hot-swappable SCSI devices
- Optional redundant hot-plug power
- Two RIO-2 and two SPCN ports

Note: The 7311 Model D20 I/O drawer initial order, or an existing 7311 Model D20 I/O drawer that is migrated from another pSeries system, must have the RIO-2 ports available (FC 6417).

7311 Model D20 I/O drawer physical package

The I/O drawer has the following physical characteristics:

- Width: 482 mm (19.0 in.)
- Depth: 610 mm (24.0 in.)
- Height: 178 mm (7.0 in.)
- Weight: 45.9 kg (101 lb.)

Figure 1-2 shows the different views of the 7311-D20 I/O drawer.

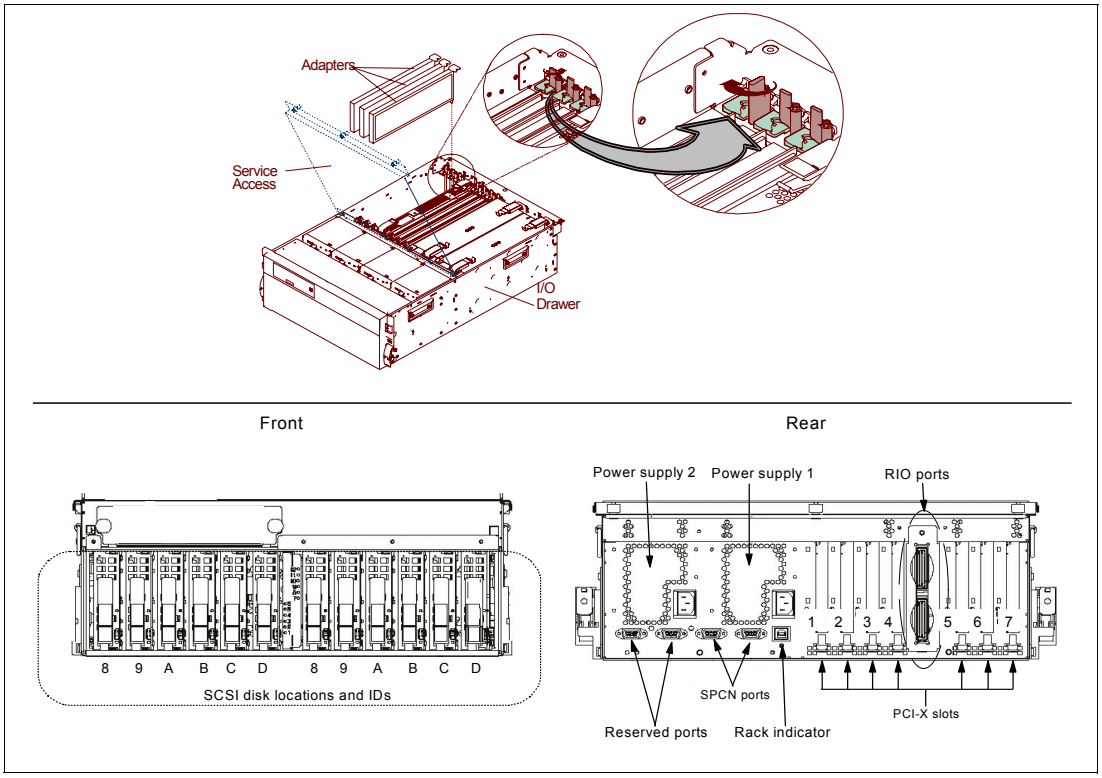


Figure 1-2 7311-D20 I/O drawer views

Note: The 7311 Model D10, and the 7311 Model D11, or the 7311 Model D20 I/O drawers are designed to be installed by an IBM service representative.

7314 Model G30 PCI-X I/O drawer

The 7314 Model G30 I/O Drawer is a rack-mountable expansion cabinet that can be attached to selected IBM System p host servers with IBM POWER6 technology. It is a half-rack width drawer that allows up to two G30 drawers to fit side-by-side in enclosure FC 7314 in the same 4 EIA units of vertical space in a 19-inch rack. Each Model G30 Drawer gives you six full-length PCI-X, 64-bit, 3.3V, PCI-X DDR adapter slots that can run at speeds up to 266 MHz.

The 7314 Model G30 I/O drawer offers a modular growth path for selected POWER6 systems. It attaches to the host system using IBMs12X Channel Interface technology. The Dual-Port 12X Channel Attach Adapters available for the Model G30 allow higher-speed data transfer rates for remote I/O drawers. A single 12X Channel I/O loop can support up to four G30 I/O drawers.

When fully configured, the system supports up to 32 Model G30 I/O Drawers attached to GX adapters (FC 1802 GX Dual Port - 12X Channel Attach) available for the GX+ slots. The combined system supports up to 200 PCI-X adapters and 12 PCIe adapters.

The I/O drawer has the following attributes:

- 4 EIA unit rack-mount enclosure (FC 7314) holding one or two G30 drawers.
- Six PCI-X DDR slots: 64-bit, 3.3V, 266 MHz. Blind-swap.
- Redundant hot-swappable power and cooling units.

- Dual-Port 12X Channel Interface adapter options:
  - short run: cables between this adapter and a host system may not exceed 3.0 Meters in length. Cables between two I/O drawers may not exceed 1.5 Meters if both I/O drawers include this short run adapter. Cables between two I/O drawers may not exceed 3.0 Meters if either of the I/O drawers includes this short run adapter.
  - long run: this adapter includes the repeater function and can support longer cable loops allowing drawers to be located in adjacent racks. 12X Cables up to 8 Meters in length can be attached to this adapter. The required 12X Cables are ordered under a separate feature number.
- Six blind-swap cassettes.

The I/O drawer physical characteristics are shown in Table 1-6.

Table 1-6 7314 G30 I/O Drawer specifications

Dimension	One G30 drawer	Mounting enclosure
Height	172 mm (6.8 in.)	176 mm (6.9 in.)
Width	224 mm (8.8 in.)	473 mm (18.6 in.)
Depth	800 mm (31.5 in.)	800 mm (31.5 in.)
Weight	20 kg (44 lb.)	45.9 kg (101 lb.) max with 2 G30 drawers

Note: 12X Channel I/O drawers cannot be mixed in a single I/O loop with RIO-2 drawers. A host system can support both RIO-2 and 12X Channel data transfer loops as long as the system supports both technologies and has the capability to support two or more independent remote I/O loops. See 2.10.6, “7314 Model G30 I/O drawer” on page 54 and 2.10.5, “7311 I/O drawer and SPCN cabling” on page 54 for more information.

I/O drawers and usable PCI slot

The different I/O drawer model types can be intermixed on a single server within the appropriate I/O loop. Depending on the system configuration, the maximum number of I/O drawers supported is different. If both 7311 and 7314 drawers are being used, the total number of I/O drawers allowed will be the values shown for the 7314-G30, assuming enough GX slots are available to configure the required RIO-2 and 12x channel adapters. For either attachment technology, up to four I/O drawers are supported in a loop.

Table 1-7 summarizes the maximum number of I/O drawers supported and the total number of PCI slots available when expansion consists of a single drawer type.

Table 1-7 Maximum number of I/O drawers supported and total number of PCI slots

System drawers/cores	Max RIO-2 drawers[a]	Max 12X Ch drawers[a]	Total number of slots					
			D11		D20		G30	
			PCI-X	PCIe	PCI-X	PCIe	PCI-X	PCIe
1 drawer / 2-core	4	4	26	4	30	4	28	4
1 drawer / 4-core	8	8	50	3[b]	58	3[b]	50	3[c]
2 drawers / 8-core	12	16	76	7[b]	88	7[b]	100	6[c]

System drawers/cores	Max RIO-2 drawers[a]	Max 12X Ch drawers[a]	Total number of slots					
			D11		D20		G30	
			PCI-X	PCIe	PCI-X	PCIe	PCI-X	PCIe
3 drawers / 12-core	16	24	102	11[b]	118	11[b]	150	9[c]
4 drawers / 16-core	20	32	128	15[b]	148	15[b]	200	12[c]

a. Up to four I/O drawers are supported in a loop
b. One PCIe slot is reserved for the Remote I/O expansion card
c. One PCIe slot per CEC drawer is reserved for the 12X channel attach expansion card.

1.3.5 Hardware Management Console models

The Hardware Management Console (HMC) is required for this system. It provides a set of functions that are necessary to manage the system, including Logical Partitioning, Capacity on Demand, inventory and microcode management, and remote power control functions.

Connection of an HMC disables the two integrated system ports.

Table 1-8 lists the HMC models available for POWER6 based systems at the time of writing. They are preloaded with the required Licensed Machine Code Version 7 (FC 0962) to support POWER6 systems, in addition to POWER5 and POWER5+ systems.

Existing HMC models 7310 can be upgraded to Licensed Machine Code Version 7 to support environments that may include POWER5, POWER5+, and POWER6 processor-based servers. Version 7 is not available for the 7315 HMCs. Licensed Machine Code Version 6 (FC 0961) is not available for 7042 HMCs, and Licensed Machine Code Version 7 (FC 0962) is not available on new 7310 HMC orders.

Table 1-8 POWER6 HMC models available

Type-model	Description
7042-C06	IBM 7042 Model C06 desktop Hardware Management Console
7042-CR4	IBM 7042 Model CR4 rack-mount Hardware Management Console

Note: POWER5 and POWER5+ processor-based servers must have firmware SF240 or later installed before being managed by a 7042 HMC or 7310 HMC with FC 0962 installed.

1.4 System racks

The system is designed to be installed in a 7014-T00 or -T42 rack. The 7014 Model T00 and T42 are 19-inch racks for general use with IBM System p rack-mount servers. An existing T00 or T42 rack can be used if sufficient space and power are available. The system is not supported in the 7014-S25 or the S11.

Note: The B42 rack is also supported.

FC 0469 Customer Specified Rack Placement provides the client the ability to specify the physical location of the system modules and attached expansion modules (drawers) in the

racks. The client's input is collected and verified through the marketing configurator (eConfig). The client's request is reviewed by eConfig for safe handling by checking the weight distribution within the rack. The manufacturing plant provides the final approval for the configuration. This information is then used by IBM Manufacturing to assemble the system components (drawers) in the rack according to the client's request.

If a system is to be installed in a non-IBM rack or cabinet, it must be ensured that the rack conforms to the EIA[2] standard EIA-310-D (see 1.4.7, "OEM rack" on page 19).

Note: It is the client's responsibility to ensure that the installation of the drawer in the preferred rack or cabinet results in a configuration that is stable, serviceable, safe, and compatible with the drawer requirements for power, cooling, cable management, weight, and rail security.

1.4.1 IBM 7014 Model T00 rack

The 1.8-meter (71-in.) Model T00 is compatible with past and present IBM System p systems. The T00 rack has the following features:

- 36 EIA units (36 U) of usable space.
- Optional removable side panels.
- Optional highly perforated front door.
- Optional side-to-side mounting hardware for joining multiple racks.
- Standard business black or optional white color in OEM format.
- Increased power distribution and weight capacity.
- Optional reinforced (ruggedized) rack feature (FC 6080) provides added earthquake protection with modular rear brace, concrete floor bolt-down hardware, and bolt-in steel front filler panels.
- Support for both AC and DC configurations.
- The rack height is increased to 1926 mm (75.8 in.) if a power distribution panel is fixed to the top of the rack.
- Up to four power distribution units (PDUs) can be mounted in the PDU bays (see Figure 1-3 on page 15), but others can fit inside the rack. See 1.4.3, "The AC power distribution unit and rack content" on page 14.
- An optional rack status beacon (FC 4690). This beacon is designed to be placed on top of a rack and cabled to servers, such as a Power 570 and other components inside the rack. Servers can be programmed to illuminate the beacon in response to a detected problem or changes in the system status.
- A rack status beacon junction box (FC 4693) should be used to connect multiple servers to the beacon. This feature provides six input connectors and one output connector for the rack. To connect the servers or other components to the junction box or the junction box to the rack, status beacon cables (FC 4691) are necessary. Multiple junction boxes can be linked together in a series using daisy chain cables (FC 4692).
- Weights:
  - T00 base empty rack: 244 kg (535 lb.)
  - T00 full rack: 816 kg (1795 lb.)

[2] Electronic Industries Alliance (EIA). Accredited by American National Standards Institute (ANSI), EIA provides a forum for industry to develop standards and publications throughout the electronics and high-tech industries.

1.4.2 IBM 7014 Model T42 rack

The 2.0-meter (79.3-inch) Model T42 addresses the client requirement for a tall enclosure to house the maximum amount of equipment in the smallest possible floor space. The features that differ in the Model T42 rack from the Model T00 include:

- 42 EIA units (42 U) of usable space (6 U of additional space).
- The Model T42 supports AC only.
- Weights:
  - T42 base empty rack: 261 kg (575 lb.)
  - T42 full rack: 930 kg (2045 lb.)

Optional Rear Door Heat eXchanger (FC 6858)

Improved cooling from the Rear Door Heat eXchanger enables clients to more densely populate individual racks, freeing valuable floor space without the need to purchase additional air conditioning units. The Rear Door Heat eXchanger features:

- Water-cooled heat exchanger door designed to dissipate heat generated from the back of computer systems before it enters the room.
- An easy-to-mount rear door design that attaches to client-supplied water, using industry standard fittings and couplings.
- Up to 15 KW (approximately 50,000 BTUs/hr) of heat removed from air exiting the back of a fully populated rack.
- One year limited warranty.

Physical specifications

The following are the general physical specifications

Approximate height	1945.5 mm (76.6 in.)
Approximate width	635.8 mm (25.03 in.)
Approximate depth back door only	1042.0 mm (41.0 in.)
Approximate depth back door and front	1098.0 mm (43.3 in.)
Approximate depth sculptured style front door	1147.0 mm (45.2 in.)
Approximate weight	31.9 kg (70.0 lb.)

Client responsibilities

Clients must ensure the following:

- Secondary water loop (to building chilled water)
- Pump solution (for secondary loop)
- Delivery solution (hoses and piping)
- Connections: Standard 3/4-inch internal threads

1.4.3 The AC power distribution unit and rack content

For rack models T00 and T42, 12-outlet PDUs are available. These include PDUs Universal UTG0247 Connector (FC 9188 and FC 7188) and Intelligent PDU+ Universal UTG0247 Connector (FC 5889 and FC 7109).

Four PDUs can be mounted vertically in the back of the T00 and T42 racks. See Figure 1-3 for the placement of the four vertically mounted PDUs. In the rear of the rack, two additional

PDUs can be installed horizontally in the T00 rack and three in the T42 rack. The four vertical mounting locations will be filled first in the T00 and T42 racks. Mounting PDUs horizontally consumes 1 U per PDU and reduces the space available for other racked components. When mounting PDUs horizontally, we recommend that you use fillers in the EIA units occupied by these PDUs to facilitate proper air-flow and ventilation in the rack.

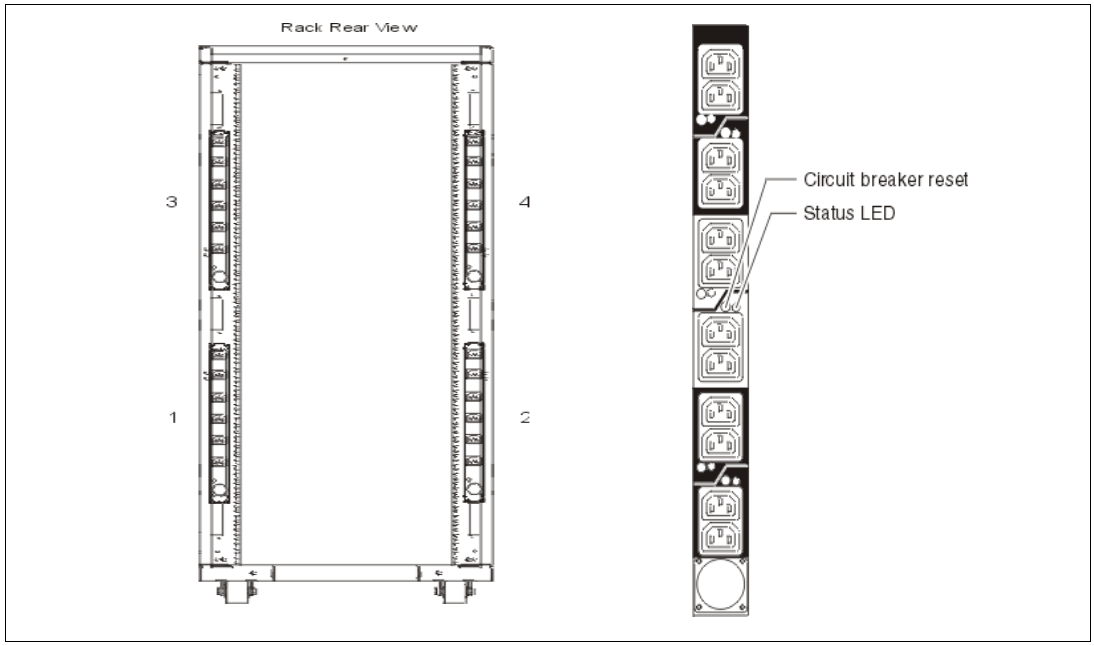


Figure 1-3 PDU placement and PDU view

For detailed power cord requirements and power cord feature codes, see *IBM System p5, eServer p5 and i5, and OpenPower Planning*, SA38-0508. For an online copy, see the IBM Systems Hardware Information Center. You can find it at:

`http://publib.boulder.ibm.com/eserver/`

Note: Ensure that the appropriate power cord feature is configured to support the power being supplied.

The Base/Side Mount Universal PDU (FC 9188) and the optional, additional, Universal PDU (FC 7188) and the Intelligent PDU+ options (FC 5889 and FC 7109) support a wide range of country requirements and electrical power specifications. The PDU receives power through a UTG0247 power line connector. Each PDU requires one PDU-to-wall power cord. Various power cord features are available for different countries and applications by varying the PDU-to-wall power cord, which must be ordered separately. Each power cord provides the unique design characteristics for the specific power requirements. To match new power requirements and save previous investments, these power cords can be requested with an initial order of the rack or with a later upgrade of the rack features.

The PDU has 12 client-usable IEC 320-C13 outlets. There are six groups of two outlets fed by six circuit breakers. Each outlet is rated up to 10 amps, but each group of two outlets is fed from one 15 amp circuit breaker.

Note: Based on the power cord that is used, the PDU can supply from 4.8 kVA to 19.2 kVA. The total kilovolt ampere (kVA) of all the drawers plugged into the PDU must not exceed the power cord limitation.

The Universal PDUs are compatible with previous models.

Note: Each system drawer to be mounted in the rack requires two power cords, which are not included in the base order. For maximum availability it is highly recommended to connect power cords from the same system to two separate PDUs in the rack. And to connect each PDU to independent power sources.

1.4.4 Intelligent Power Distribution Unit (iPDU)

Energy consumption is becoming a large issue in computer-based businesses. The energy required to power and cool computers can be a significant cost to a business – reducing profit margins and consuming resources.

For all systems without an internal thermal and power consumption method the IBM Intelligent Power Distribution Management (IPDU) provides a solution to measure and collect power data. An iPDU (FC 5889) mounts in a rack and provides power outlets for the servers to plug into.

The following list shows the characteristics of an iPDU:

Input connector	Connect power cord to this connector
Power outlets	Power outlet for devices. There are nine or 12 power outlets, depending on the model
RS232 serial connector	Update firmware
RJ45 console connector	Provides a connection using a DB9-to-RJ45 provided cable to a notebook computer as a configuration console.
RJ45 Ethernet (LAN) connector	Port to configure the iPDU through a LAN. Speed is 10/100 auto sensed.

When a configured iPDU is selected the following dialog panel (as in IBM Systems Director) will appear as shown in Figure 1-4 on page 17.

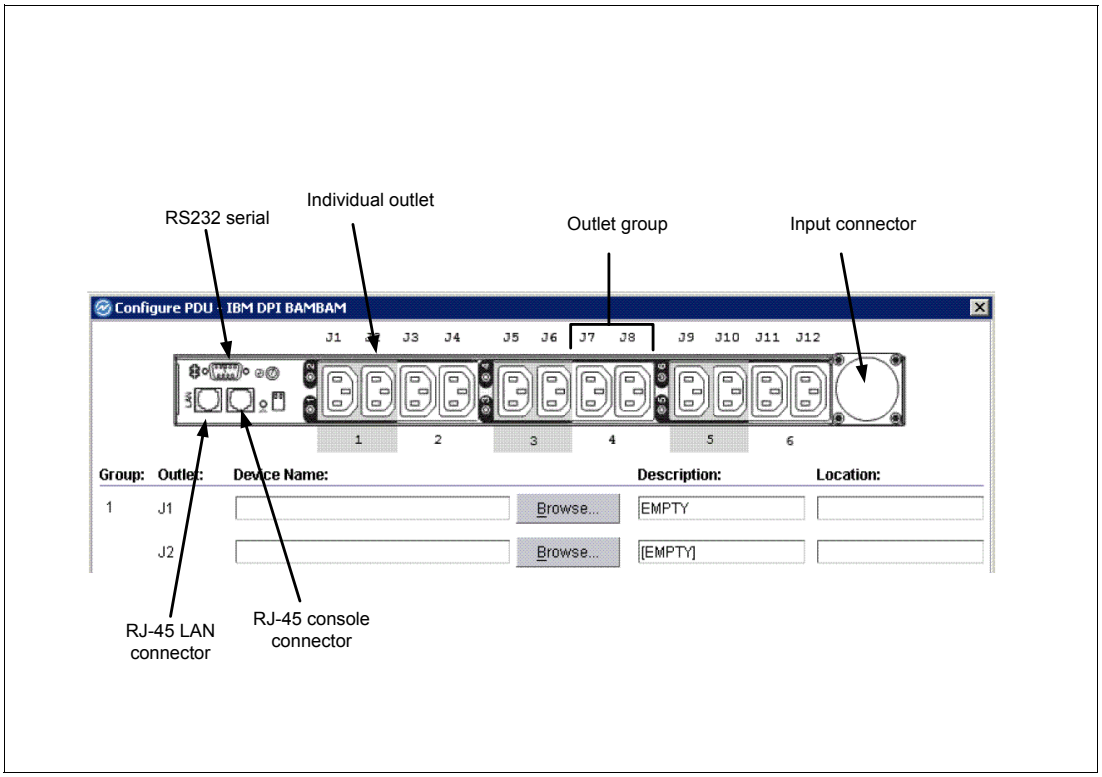

RS232 serial
Individual outlet
Outlet group
Input connector
Configure PDU - IBM DPI BAMBAM
J1 J2 J3 J4 J5 J6 J7 J8 J9 J10 J11 J12
1 2 3 4 5 6
Group: Outlet: Device Name: Description: Location:
1 J1 Browse... EMPTY
J2 Browse... [EMPTY]
RJ-45 LAN connector
RJ-45 console connector

Figure 1-4 Intelligent Power Distribution Unit

In this panel, outlet names and outlet group names are shown. Each iPDU node will contain either node group outlets or individual outlets. For further information of integration and IBM Director functionality, see:

http://www.ibm.com/systems/management/director/about

1.4.5 Rack-mounting rules

The system consists of one to four CEC enclosures. Each enclosure occupies 4U of vertical rack space. The primary considerations that should be accounted for when mounting the system into a rack are:

- The Power 570 is designed to be placed at any location in the rack. For rack stability, it is advisable to start filling a rack from the bottom.
- For configurations with two, three, or four drawers, all drawers must be installed together in the same rack, in a contiguous space of 8 U, 12 U, or 16 U within the rack. The uppermost enclosure in the system is the base enclosure. This enclosure will contain the primary active Service Processor and the Operator Panel.
- Any remaining space in the rack can be used to install other systems or peripherals, provided that the maximum permissible weight of the rack is not exceeded and the installation rules for these devices are followed.
- The 7014-T42 rack is constructed with a small flange at the bottom of EIA location 37. When a system is installed near the top of a 7014-T42 rack, no system drawer can be installed in EIA positions 34, 35, or 36. This is to avoid interference with the front bezel or with the front flex cable, depending on the system configuration. A two-drawer system cannot be installed above position 29. A three-drawer system cannot be installed above position 25. A four-drawer system cannot be installed above position 21. (The position number refers to the bottom of the lowest drawer.)

- When a system is installed in an 7014-T00 or -T42 rack that has no front door, a Thin Profile Front Trim Kit must be ordered for the rack. The required trim kit for the 7014-T00 rack is FC 6246. The required trim kit for the 7014-T42 rack is FC 6247.
- The design of the 570 is optimized for use in a 7014-T00 or -T42 rack. Both the front cover and the processor flex cables occupy space on the front left side of an IBM 7014 rack that may not be available in typical non-IBM racks.
- Acoustic Door features are available with the 7014-T00 and 7014-T42 racks to meet the lower acoustic levels identified in the specification section of this document. The Acoustic Door feature can be ordered on new T00 and T42 racks or ordered for the T00 and T42 racks that clients already own.

1.4.6 Useful rack additions

This section highlights useful additions to a rack.

IBM 7214 Model 1U2 SAS Storage Enclosure

IBM 7212 Model 102 IBM TotalStorage storage device enclosure

The IBM 7212 Model 102 is designed to provide efficient and convenient storage expansion capabilities for selected System p servers. The IBM 7212 Model 102 is a 1 U rack-mountable option to be installed in a standard 19-inch rack using an optional rack-mount hardware feature kit. The 7212 Model 102 has two bays that can accommodate any of the following storage drive features:

- A Digital Data Storage (DDS) Gen 5 DAT72 Tape Drive provides a physical storage capacity of 36 GB (72 GB with 2:1 compression) per data cartridge.
- A VXA-2 Tape Drive provides a media capacity of up to 80 GB (160 GB with 2:1 compression) physical data storage capacity per cartridge.
- A Digital Data Storage (DDS-4) tape drive provides 20 GB native data capacity per tape cartridge and a native physical data transfer rate of up to 3 MBps that uses a 2:1 compression so that a single tape cartridge can store up to 40 GB of data.
- A DVD-ROM drive is a 5 1/4-inch, half-high device. It can read 640 MB CD-ROM and 4.7 GB DVD-RAM media. It can be used for alternate IPL[3] (IBM-distributed CD-ROM media only) and program distribution.
- A DVD-RAM drive with up to 2.7 MBps throughput. Using 3:1 compression, a single disk can store up to 28 GB of data. Supported DVD disk native capacities on a single DVD-RAM disk are as follows: up to 2.6 GB, 4.7 GB, 5.2 GB, and 9.4 GB.

Flat panel display options

The IBM 7316-TF3 Flat Panel Console Kit can be installed in the system rack. This 1 U console uses a 15-inch thin film transistor (TFT) LCD with a viewable area of 304.1 mm x 228.1 mm and a 1024 x 768 pels[4] resolution. The 7316-TF3 Flat Panel Console Kit has the following attributes:

- Flat panel color monitor
- Rack tray for keyboard, monitor, and optional VGA switch with mounting brackets
- IBM Travel Keyboard mounts in the rack keyboard tray (Integrated Trackpoint and UltraNav)

[3] Initial program load
[4] Picture elements

IBM PS/2 Travel Keyboards are supported on the 7316-TF3 for use in configurations where only PS/2 keyboard ports are available.

The IBM 7316-TF3 Flat Panel Console Kit provides an option for the USB Travel Keyboards with UltraNav. The keyboard enables the 7316-TF3 to be connected to systems that do not have PS/2 keyboard ports. The USB Travel Keyboard can be directly attached to an available integrated USB port or a supported USB adapter (FC 2738) on System p5 servers or 7310-CR3 and 7315-CR3 HMCs.

The Netbay LCM (Keyboard/Video/Mouse) Switch (FC 4202) provides users single-point access and control of up to 64 servers from a single console. The Netbay LCM Switch has a maximum video resolution of 1600 x 280 and mounts in a 1 U drawer behind the 7316-TF3 monitor. A minimum of one LCM feature (FC 4268) or USB feature (FC 4269) is required with a Netbay LCM Switch (FC 4202). Each feature can support up to four systems. When connecting to a Power 570, FC 4269 provides connection to the server USB ports.

When selecting the LCM Switch, consider the following information:

- The KVM Conversion Option (KCO) cable (FC 4268) is used with systems with PS/2 style keyboard, display, and mouse ports.
- The USB cable (FC 4269) is used with systems with USB keyboard or mouse ports.
- The switch offers four ports for server connections. Each port in the switch can connect a maximum of 16 systems:
  - One KCO cable (FC 4268) or USB cable (FC 4269) is required for every four systems supported on the switch.
  - A maximum of 16 KCO cables or USB cables per port can be used with the Netbay LCM Switch to connect up to 64 servers.

Note: A server microcode update might be required on installed systems for boot-time System Management Services (SMS) menu support of the USB keyboards. The update might also be required for the LCM switch on the 7316-TF3 console (FC 4202). For microcode updates, see the following URL:

`http://techsupport.services.ibm.com/server/mdownload`

We recommend that you have the 7316-TF3 installed between EIA 20 to 25 of the rack for ease of use. The 7316-TF3 or any other graphics monitor requires the POWER GXT135P graphics accelerator (FC 1980) to be installed in the server, or some other graphics accelerator, if supported.

1.4.7 OEM rack

The system can be installed in a suitable OEM rack, provided that the rack conforms to the EIA-310-D standard for 19-inch racks. This standard is published by the Electrical Industries Alliance, and a summary of this standard is available in the publication *IBM System p5, eServer p5 and i5, and OpenPower Planning*, SA38-0508.

The key points mentioned in this documentation are as follows:

- The front rack opening must be 451 mm wide + 0.75 mm (17.75 in. + 0.03 in.), and the rail-mounting holes must be 465 mm + 0.8 mm (18.3 in. + 0.03 in.) apart on center (horizontal width between the vertical columns of holes on the two front-mounting flanges and on the two rear-mounting flanges). See Figure 1-5 on page 20 for a top view showing the specification dimensions.

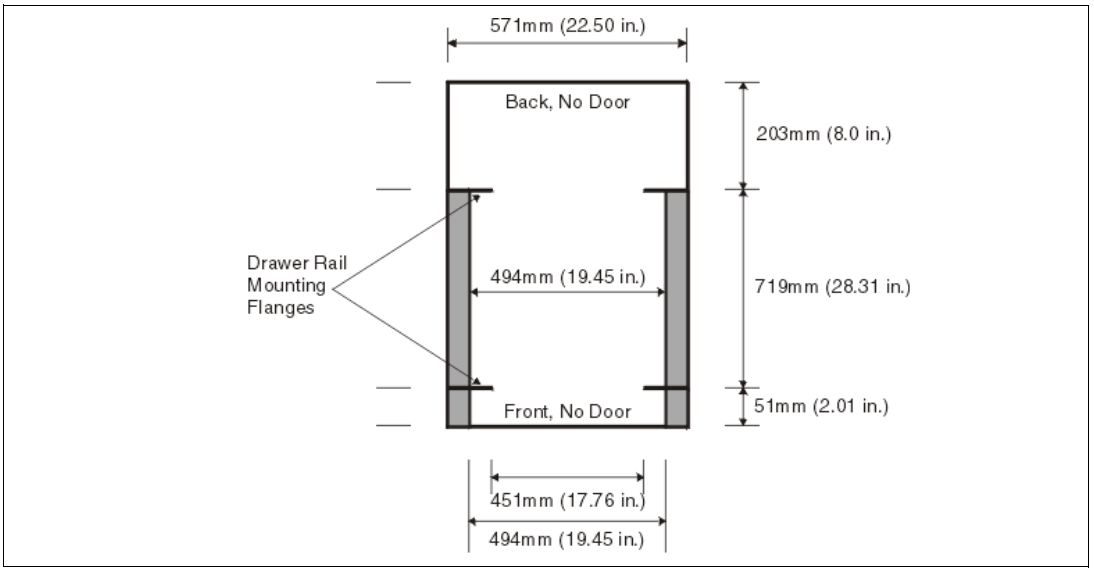


Figure 1-5 Top view of non-IBM rack specification dimensions

- The vertical distance between the mounting holes must consist of sets of three holes spaced (from bottom to top) 15.9 mm (0.625 in.), 15.9 mm (0.625 in.), and 12.67 mm (0.5 in.) on center, making each three-hole set of vertical hole spacing 44.45 mm (1.75 in.) apart on center. Rail-mounting holes must be 7.1 mm + 0.1 mm (0.28 in. + 0.004 in.) in diameter. See Figure 1-6 and Figure 1-7 on page 21 for the top and bottom front specification dimensions.

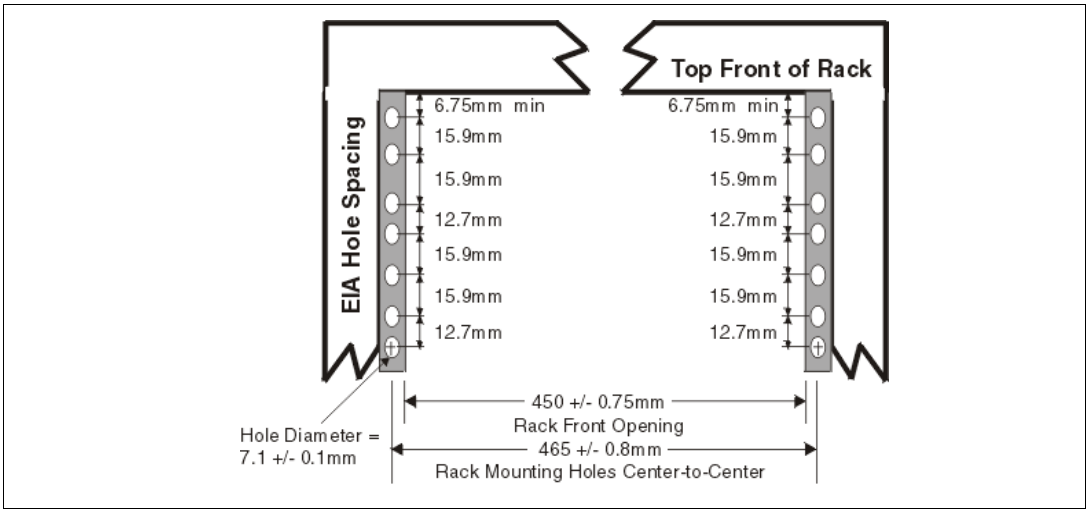


Figure 1-6 Rack specification dimensions, top front view

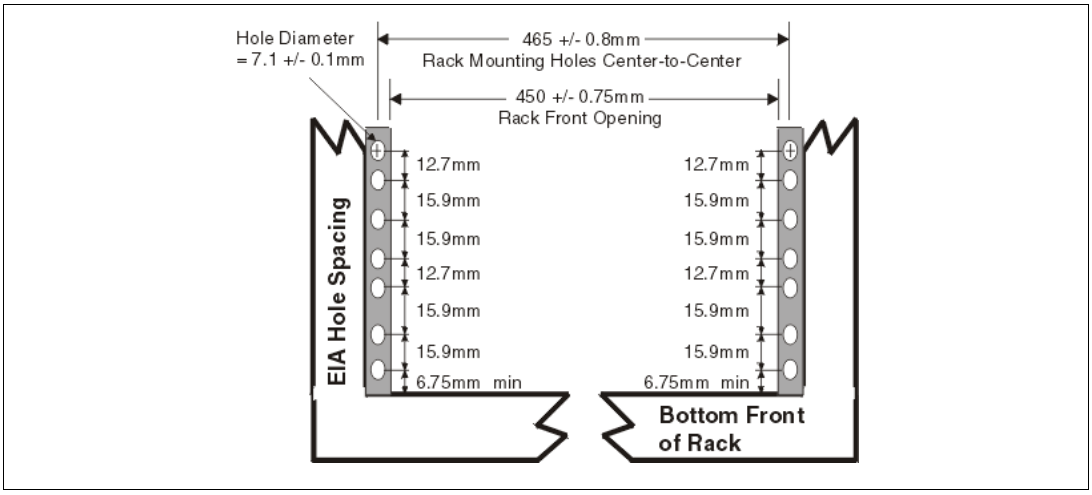


Figure 1-7 Rack specification dimensions, bottom front view

- It might be necessary to supply additional hardware, such as fasteners, for use in some manufacturer's racks.
- The system rack or cabinet must be capable of supporting an average load of 15.9 kg (35 lb.) of product weight per EIA unit.
- The system rack or cabinet must be compatible with drawer mounting rails, including a secure and snug fit of the rail-mounting pins and screws into the rack or cabinet rail support hole.

Note: The OEM rack must only support ac-powered drawers. We strongly recommend that you use a power distribution unit (PDU) that meets the same specifications as the PDUs to supply rack power. Rack or cabinet power distribution devices must meet the drawer power requirements, as well as the requirements of any additional products that will be connected to the same power distribution device.

2

Architecture and technical overview

This chapter discusses the overall system architecture represented by Figure 2-1, with its major components described in the following sections. The bandwidths that are provided throughout the section are theoretical maximums used for reference. You should always obtain real-world performance measurements using production workloads.

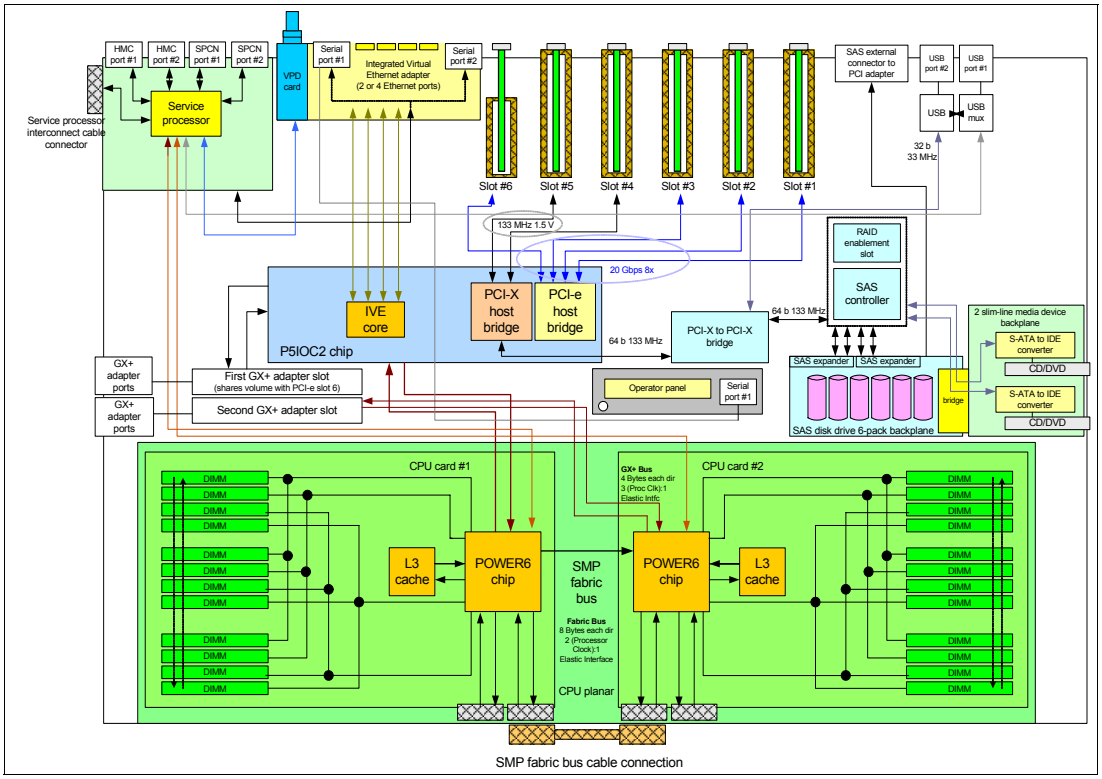


Figure 2-1 570 logic data flow

© Copyright IBM Corp. 2008. All rights reserved.

2.1 The POWER6 processor

The POWER6 processor capitalizes on all the enhancements brought by the POWER5 chip.

Two of the enhancements of the POWER6 processor is the ability to do processor instruction retry and alternate processor recovery. This significantly reduces exposure to both hard (logic) and soft (transient) errors in the processor core.

Processor instruction retry

Soft failures in the processor core are transient errors. When an error is encountered in the core, the POWER6 processor will first automatically retry the instruction. If the source of the error was truly transient, the instruction will succeed and the system will continue as before. On predecessor IBM systems, this error would have caused a checkstop.

Alternate processor retry

Hard failures are more difficult, being true logical errors that will be replicated each time the instruction is repeated. Retrying the instruction will not help in this situation because the instruction will continue to fail. Systems with POWER6 processors introduce the ability to extract the failing instruction from the faulty core and retry it elsewhere in the system, after which the failing core is dynamically deconfigured and called out for replacement. The entire process is transparent to the partition owning the failing instruction. Systems with POWER6 processors are designed to avoid what would have been a full system outage.

POWER6 single processor checkstopping

Another major advancement in POWER6 processors is single processor checkstopping. A processor checkstop would result in a system checkstop. A new feature in the 570 is the ability to contain most processor checkstops to the partition that was using the processor at the time. This significantly reduces the probability of any one processor affecting total system availability.

POWER6 cache availability

In the event that an uncorrectable error occurs in L2 or L3 cache, the system will be able to dynamically remove the offending line of cache without requiring a reboot. In addition POWER6 utilizes a L1/L2 cache design and a write-through cache policy on all levels, helping to ensure that data is written to main memory as soon as possible.

Figure 2-2 on page 25 shows a high-level view of the POWER6 processor.

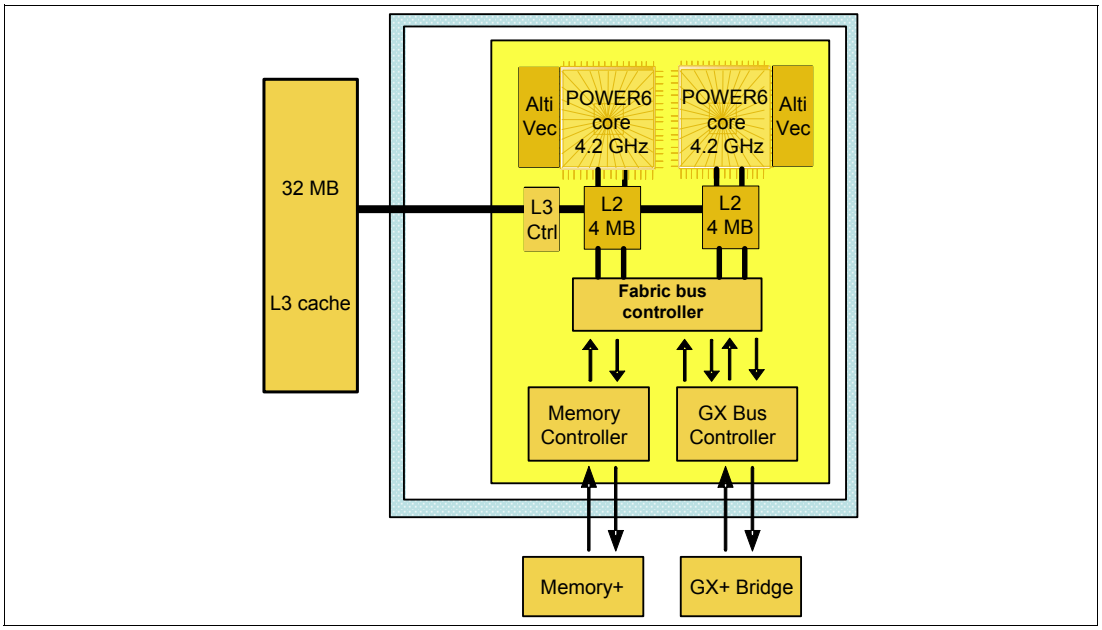


Figure 2-2 POWER6 processor

The CMOS 11S0 lithography technology in the POWER6 processor uses a 65 nm fabrication process, which enables:

- Performance gains through faster clock rates from 3.5 GHz, 4.2 GHz up to 4.7 GHz.
- Physical size of 341 mm.

The POWER6 processor consumes less power and requires less cooling. Thus, you can use the POWER6 processor in servers where previously you could only use lower frequency chips due to cooling restrictions.

The 64-bit implementation of the POWER6 design provides the following additional enhancements:

- Compatibility of 64-bit architecture
  - Binary compatibility for all POWER and PowerPC® application code level
  - Support of partition migration
  - Support of virtualized partition memory
  - Support of four page sizes : 4 KB, 64 KB, 16 MB, and 16 GB
- High frequency optimization
  - Designed to operate at maximum speed of 5 GHz
- Superscalar core organization
  - Simultaneous Multithreading: two threads
- In order dispatch of five operations (single thread), seven operations (Simultaneous Multithreading) to nine execution units:
    - Two load or store operations
    - Two fixed-point register-register operations
    - Two floating-point operations
    - One branch operation

The POWER6 processor implements the 64-bit IBM Power Architecture® technology. Each POWER6 chip incorporates two dual-threaded Simultaneous Multithreading processor cores, a private 4 MB level 2 cache (L2) for each processor, a 36 MB L3 cache controller shared by the two processors, integrated memory controller and data interconnect switch and support logic for dynamic power management, dynamic configuration and recovery, and system monitoring.

2.1.1 Decimal floating point

This section describes the behavior of the decimal floating-point processor, the supported data types, formats, and classes, and the usage of registers.

The decimal floating-point (DFP) processor shares the 32 floating-point registers (FPRs) and the floating-point status and control register (FPSCR) with the binary floating-point (BFP) processor. However, the interpretation of data formats in the FPRs, and the meaning of some control and status bits in the FPSCR are different between the BFP and DFP processors.

The DFP processor supports three DFP data formats:

- DFP32 (single precision)
- DFP64 (double precision)
- DFP128 (quad precision)

Most operations are performed on the DFP64 or DFP128 format directly. Support for DFP32 is limited to conversion to and from DFP64. For some operations, the DFP processor also supports operands in other data types, including signed or unsigned binary fixed-point data, and signed or unsigned decimal data.

DFP instructions are provided to perform arithmetic, compare, test, quantum-adjustment, conversion, and format operations on operands held in FPRs or FPR pairs.

Arithmetic instructions These instructions perform addition, subtraction, multiplication, and division operations.

Compare instructions These instructions perform a comparison operation on the numerical value of two DFP operands.

Test instructions These instructions test the data class, the data group, the exponent, or the number of significant digits of a DFP operand.

Quantum-adjustment instructions These instructions convert a DFP number to a result in the form that has the designated exponent, which may be explicitly or implicitly specified.

Conversion instructions These instructions perform conversion between different data formats or data types.

Format instructions These instructions facilitate composing or decomposing a DFP operand.

For example, SAP NetWeaver® 7.10 ABAP™ kernel introduces a new SAP ABAP data type called *DECFLOAT* to enable more accurate and consistent results from decimal floating point computations. The decimal floating point (DFP) support by SAP NetWeaver leverages the built-in DFP feature of POWER6 processors. This allows for highly simplified ABAP-coding while increasing numeric accuracy and with a potential for significant performance improvements.

2.1.2 AltiVec and Single Instruction, Multiple Data

IBM Semiconductor's advanced Single Instruction, Multiple Data (SIMD) technology based on the AltiVec instruction set is designed to enable exceptional general-purpose processing power for high-performance POWER processors. This leading-edge technology is engineered to support high-bandwidth data processing and algorithmic-intensive computations, all in a single-chip solution

With its computing power, AltiVec technology also enables high-performance POWER processors to address markets and applications in which performance must be balanced with power consumption, system cost and peripheral integration.

The AltiVec technology is a well known environment for software developers who want to add efficiency and speed to their applications. A 128-bit vector execution unit was added to the architecture. This engine operates concurrently with the existing integer and floating-point units and enables highly parallel operations, up to 16 operations in a single clock cycle. By leveraging AltiVec technology, developers can optimize applications to deliver acceleration in performance-driven, high-bandwidth computing.

The AltiVec technology is not comparable to the IBM POWER6 processor implementation, using the Simultaneous Multithreading functionality.

2.2 IBM EnergyScale technology

IBM EnergyScale™ technology is featured on the IBM POWER6 processor-based systems. It provides functions to help the user understand and control IBM server power and cooling usage.

In this section we will describe IBM EnergyScale features and hardware and software requirements

Power Trending EnergyScale provides continuous power usage data collection. This enables the administrators with the information to predict power consumption across their infrastructure and to react to business and processing needs. For example, an administrator could adjust server consumption to reduce electrical costs. To collect power data for the 570 you need to power it through an Intelligent Power Distribution Unit (iPDU). Other systems that support power trending collect the information internally and do not require any additional hardware.

Power Saver Mode Power Saver Mode reduces the voltage and frequency by a fixed percentage. This percentage is predetermined to be within a safe operating limit and is not user configurable. Under current implementation this is a 14% frequency drop. When CPU utilization is low, Power Saver Mode has no impact on performance. Power Saver Mode can reduce the processor usage up to a 30%. Power Saver Mode is not supported during boot or re-boot although it is a persistent condition that will be sustained after the boot when the system starts executing instructions. Power Saver is only supported with 4.2 GHz processors and faster.

Power Capping Power Capping enforces a user specified limit on power usage. Power Capping is not a power saving mechanism. It enforces power caps by actually throttling the processor(s) in the system, degrading performance significantly. The idea of a power cap is to set something that should never be reached but frees up *margined*

power in the data center. The margined power is the amount of extra power that is allocated to a server during its installation in a datacenter. It's based on the server environmental specifications that usually are never reached. Since server specifications are always based on maximum configurations and worst case scenarios.

Processor Core Nap The IBM POWER6 processor uses a low-power mode called Nap that stops processor execution when there is no work to do on that processor core, both threads are idle. Nap mode allows the hardware to clock off most of the circuits inside the processor core. Reducing active power consumption by turning off the clocks allows the temperature to fall, which further reduces leakage (static) power of the circuits causing a cumulative effect. Unlicensed cores are kept in core Nap until they are licensed and return to core Nap whenever they are unlicensed again.

EnergyScale for I/O IBM POWER6 processor-based systems automatically power off pluggable, PCI adapter slots that are empty or not being used to save approximately 14 watts per slot. System firmware automatically scans all pluggable PCI slots at regular intervals looking for ones that meet the criteria for being not in use and powers them off. This support is available for all POWER6 processor-based servers, and the expansion units that they support. Note that it applies to hot pluggable PCI slots only.

2.2.1 Hardware and software requirements

This sections summarizes the supported systems and software user interfaces for EnergyScale functions.

Table 2-1 EnergyScale systems support

	Power trending	Power saver mode	Power capping	Processor Nap	I/O
7998-61X	Y	Y	Y	Y	N
8203-E4A	Y	Y	Y	Y	Y
8204-E8A	Y	Y	Y	Y	Y
9117-MMA (< 4.20 GHz)	Y - via iPDU[a]	N	N	Y	Y
9117-MMA (>= 4.2 GHz)	Y - via iPDU[a]	Y[b]	N	Y	Y

a. An iPDU is required for this support. The feature code 5889 is for a base iPDU in a rack while the feature code 7109 is for additional iPDUs for the same rack. Supported racks are 7014-B42, 7014-S25, 7014-T00 and 7014-T42.
b. Only supported if GX Dual Port RIO-2 Attach (FC 1800) is not present.

The primary user interface for EnergyScale features on a POWER6 based system is IBM Systems Director Active Energy Manager™ running within IBM Director

Table 2-2 on page 29 shows the ASMI, HMC and Active Energy Manager interface support.

Table 2-2 EnergyScale function's software interfaces

EnergyScale functions	ASMI	HMC	Active Energy Manager
Power Trending	N	N	Y
Power Saver Mode	Y	Y	Y
Schedule Power Saver Mode Operation	N	Y	Y
Power Capping	N	N	Y
Schedule Power Capping Operation	N	N	Y

2.3 Processor cards

In the 570, the POWER6 processors, associated L3 cache chip, and memory DIMMs are packaged in processor cards. The 570 uses a dual-core processor module for a 2-core, 4-core, 8-core, 12-core, and 16-core configuration running at 3.5 GHz, 4.2 GHz, or 4.7 GHz.

The 570 has two processor sockets on the system planar. Each socket will accept a processor card feature. A single CEC may have one or two processor cards installed. A system with two, three, or four CEC must have two processor cards in each CEC.

Each processor can address all the memory on the processor card. Access to memory behind another processor is accomplished through the fabric buses.

The 2-core 570 processor card contains a dual-core processor chip, a 36 MB L3 cache chip and the local memory storage subsystem.

Figure 2-3 shows a layout view of a 570 processor card and associated memory.

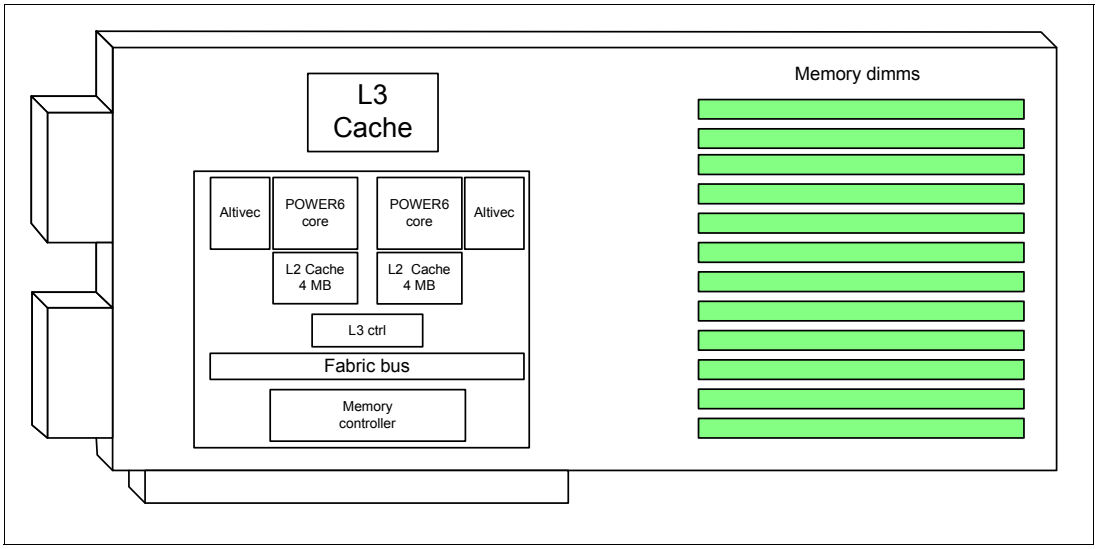


Figure 2-3 The 570 processor card with DDR2 memory socket layout view

The storage structure for the POWER6 processor is a distributed memory architecture that provides high-memory bandwidth, although each processor can address all memory and

sees a single shared memory resource. They are interfaced to 12 memory slots, where as each memory DIMM has its own memory buffer chip and are interfaced in a point-to-point connection.

I/O connects to the 570 processor module using the GX+ bus. The processor module provides a single GX+ bus. The GX+ bus provides an interface to I/O devices through the RIO-2 connections or a 12X Channel attach connections.

2.3.1 Processor drawer interconnect cables

In combined systems that are made of more than one 570 building block, the connection between processor cards in different building blocks is provided with a processor drawer interconnect cable. Different processor drawer interconnect cables are required for the different numbers of 570 building blocks that a combined system can be made of, as shown in Figure 2-4.

Because of the redundancy and fault recovery built-in to the system interconnects, a drawer failure does not represent a system failure. Once a problem is isolated and repaired, a system reboot may be required to reestablish full bus speed, if the failure was specific to the interconnects.

The SMP fabric bus that connects the processors of separate 570 building blocks is routed on the interconnect cable that is routed external to the building blocks. The flexible cable attaches directly to the processor cards, at the front of the 570 building block, and is routed behind the front covers (bezels) of the 570 building blocks. There is an optimized cable for each drawer configuration. Figure 2-4 illustrates the logical fabric bus connections between the drawers, and shows the additional space required left of the bezels for rack installation.

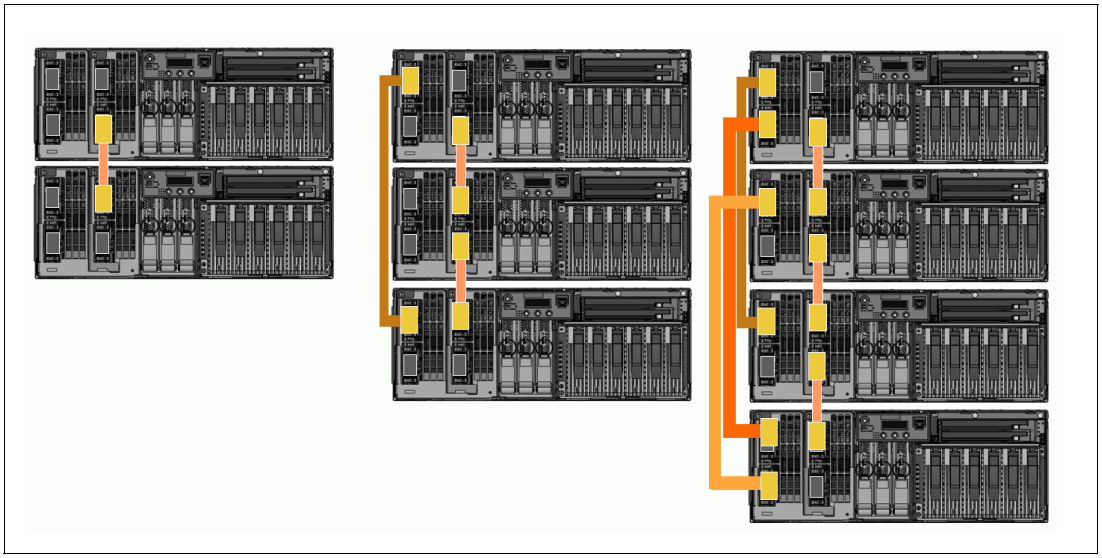


Figure 2-4 Logical 570 building block connection

2.3.2 Processor clock rate

The 570 system features base 2-core, 4-core, 8-core, 12-core, and 16-core configurations with the POWER6 processor running at 3.5 GHz, 4.2 GHz, and 4.7 GHz.

Note: Any system made of more than one processor card must have all processor cards running at the same speed.

To verify the processor characteristics on a system running at 4.2 GHz, use one of the following commands:

- **`lsattr -El proc`*X***

  Where *X* is the number of the processor, for example, proc0 is the first processor in the system. The output from the command is similar to the following output (*False*, as used in this output, signifies that the value cannot be changed through an AIX command interface):

```
frequency   4208000000          Processor Speed       False
smt_enabled true                Processor SMT enabled False
smt_threads 2                   Processor SMT threads False
state       enable              Processor state       False
type        powerPC_POWER6    Processor type        False
```

- **`pmcycles -m`**

  The **`pmcycles`** command (available with AIX) uses the performance monitor cycle counter and the processor real-time clock to measure the actual processor clock speed in MHz. The following output is from a 4-core 570 system running at 4.2 GHz with simultaneous multithreading enabled:

```
Cpu 0 runs at 4208 MHz
Cpu 1 runs at 4208 MHz
Cpu 2 runs at 4208 MHz
Cpu 3 runs at 4208 MHz
Cpu 4 runs at 4208 MHz
Cpu 5 runs at 4208 MHz
Cpu 6 runs at 4208 MHz
Cpu 7 runs at 4208 MHz
```

Note: The **`pmcycles`** command is part of the bos.pmapi fileset. Use the **`lslpp -l bos.pmapi`** command to determine if it is installed on your system.

2.4 Memory subsystem

When you consider a 570 initial order, the memory controller is internal to the POWER6 processor and it interfaces any of the memory buffer chips within the pluggable fully buffered DIMMs (12 slots available per processor card, as described in 1.3.2, “Memory features” on page 6).

2.4.1 Fully buffered DIMM

Fully buffered DIMM is a memory technology which can be used to increase reliability, speed and density of memory subsystems. Conventionally, data lines from the memory controller have to be connected to data lines in every DRAM module. As memory width, as well as access speed, increases, the signal degrades at the interface of the bus and the device. This limits the speed or the memory density. The fully buffered DIMMs take a different approach because directs signaling interface between the memory controller and the DRAM chips, splitting it into two independent signaling interfaces with a buffer between them. The interface

between the memory controller and the buffer is changed from a shared parallel interface to a point-to-point serial interface (see Figure 2-5).

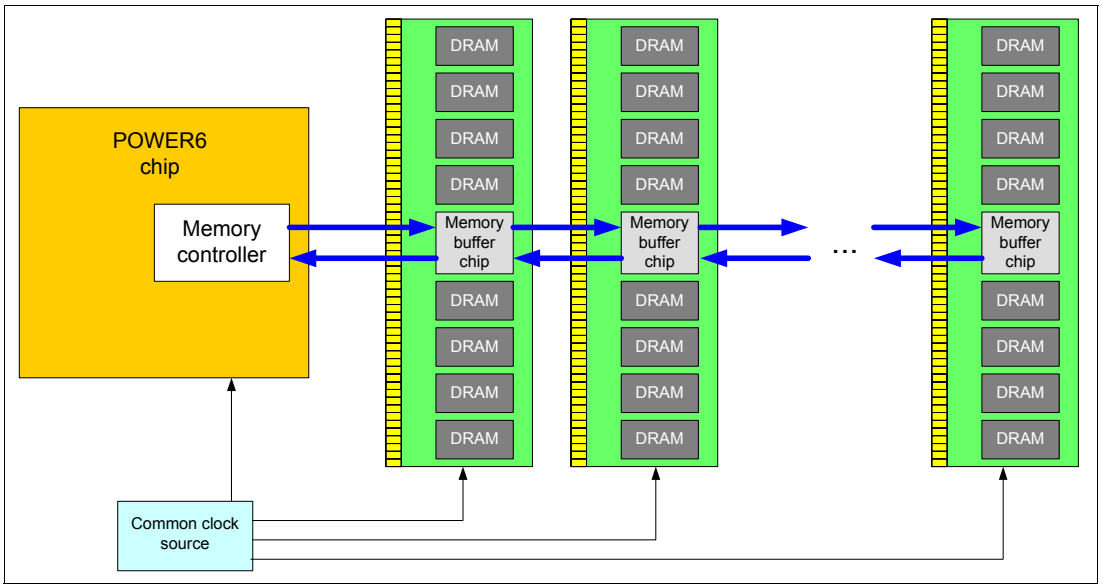


Figure 2-5 Fully buffered DIMMs architecture

The result of the fully buffered memory DIMMs implementation is an enhanced scalability and throughput.

2.4.2 Memory placements rules

The minimum memory capacity for a 570 initial order is 2 GB when a 3.5 GHz, 4.2 GHz, or 4.7 GHz system is configured with two processor-cores. FC 5620, FC 5622, and FC 7380 processor cards support up to 12 fully buffered DIMM slots and DIMMs must be installed in quads. Then the quads are organized as follows:

- First quad includes J0A, J0B, J0C, and J0D memory slots
- Second quad includes J1A, J1B, J1C, and J1D memory slots
- Third quad includes J2A, J2B, J2C, and J2D memory slots

See Figure 2-6 on page 33 to locate any available quad.

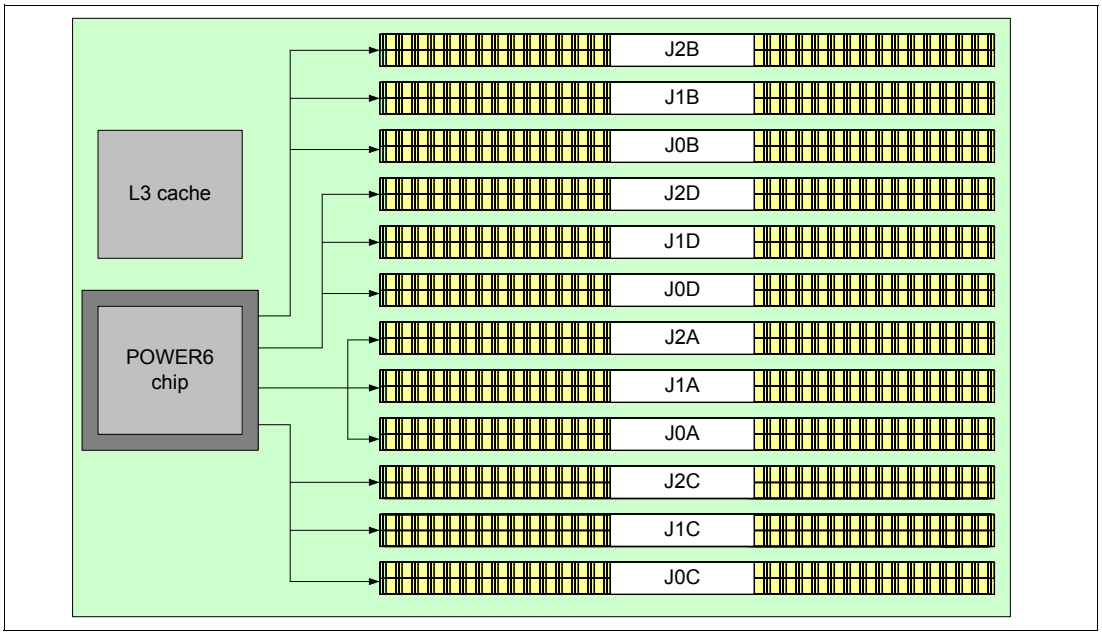


Figure 2-6 Memory DIMM slots for FC 5620, FC 5622, and FC 7380

In addition to the quad placement rules, minimum memory required depends from the number of processor-cores configured in the 570:

- 2 GB is the minimum memory required for a 2-core system
- 4 GB is the minimum memory required for a 4-core system
- 8 GB is the minimum memory required for an 8-core system
- 16 GB is the minimum memory required for a 16-core system

Every processor card in a 570 configuration requires a memory quad.

The maximum installable memory is 192 GB per any 570 drawer, thus a fully configured 570 supports up to 768 GB (48 GB per processor-core).

When configuring the memory in a 570, placing 2 memory features (8 DIMMs) on a single processor card will provide the maximum available memory bandwidth. Adding the third memory feature will provide additional memory capacity but will not increase memory bandwidth. System performance that is dependent on memory bandwidth can be improved by purchasing two smaller features per processor card as opposed to one large feature per processor card. To achieve this, when placing an order, ensure the order has 2X memory features for every processor card feature on the order.

2.4.3 Memory consideration for model migration from p5 570 to 570

A p5 570 (based on POWER5 or POWER5+ processor) can be migrated to a 570. Since the 570 supports only DDR2 memory, if the initial p5 570 server to migrate has DDR2 memory, it can be migrated to the target 570 that requires FC 5621 processor card to accept it. Additional memory can be also included in the model migration order.

In FC 5621 processor card, the memory controller interfaces to four memory buffer chips per processor card with 8 memory slots available to be populated with available DDR2 memory DIMMs migrated from the p5 570 server (see Figure 2-7 on page 34 for memory DIMM slots reference).

If the initial p5 570 server does not configure supported DDR2 memory, then the target 570 can be configured with FC 5622 processor card and the desired amount of memory must be included in the model migration order.

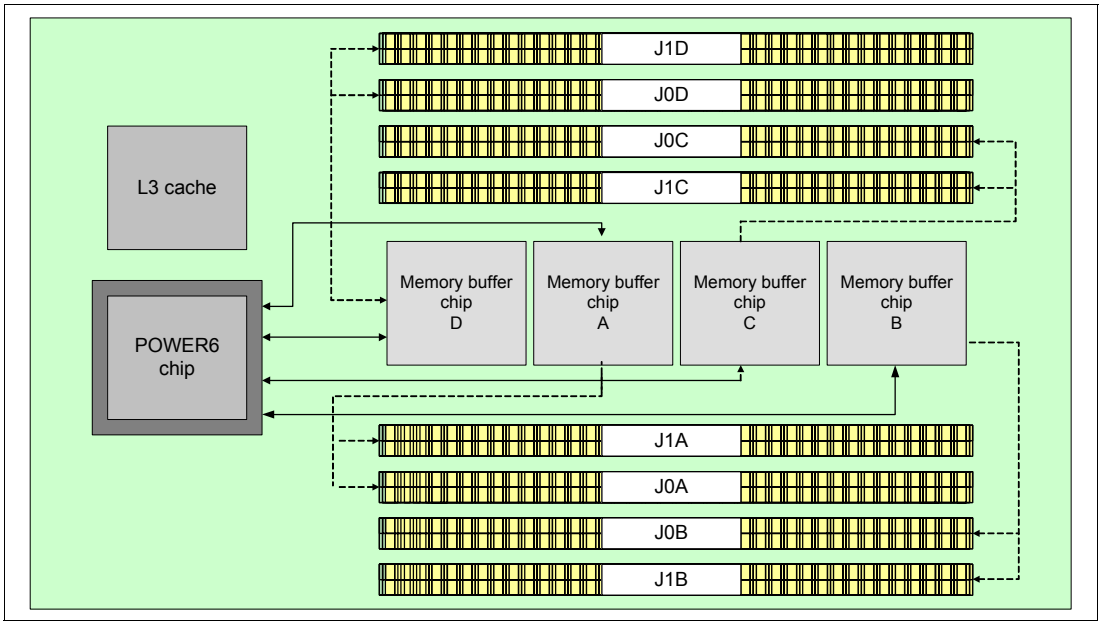


Figure 2-7 Memory DIMM slots for FC 5621

Important: The process to migrate a p5 570 to a 570 requires analysis of the existing p5 570 memory DIMMs. Contact an IBM service representative before issuing the configuration upgrade order.

A 570 with FC 5621 processor cards can be expanded by purchasing additional 570 enclosures with FC 5622 processor cards. FC 5621 and FC 5622 cannot be mixed within the same 570 enclosure but can be mixed in the same system. Maximum memory configurable depends from the number of FC 5621 and FC 5622 processor cards available in the fully combined 570 system.

2.4.4 OEM memory

OEM memory is not supported or certified for use in IBM System p servers. If the 570 is populated with OEM memory, you could experience unexpected and unpredictable behavior, especially when the system is using Micro-Partitioning technology.

All IBM memory is identified by an IBM logo and a white label that is printed with a barcode and an alphanumeric string, as illustrated in Figure 2-8 on page 35.

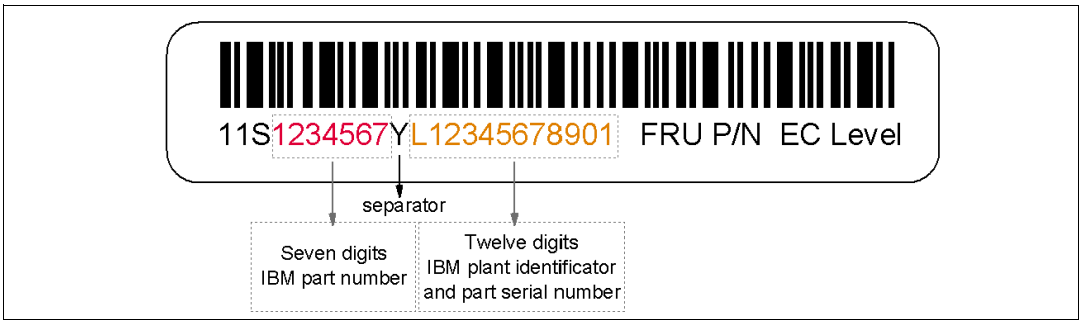


Figure 2-8 IBM memory certification label

2.4.5 Memory throughput

The memory subsystem throughput is based on the speed of the memory. On processor, there are four memory channels, each with single 2 byte read and 1 byte write. Memory channels of POWER6 memory controller are connected to Memory buffers. The processor chip has two POWER6 processors. The DDR2 bus allows double reads or writes per clock cycle. If a 667 MHz memory feature is selected, the throughput is (4 x 2 x 2 x 2 x 667) + (4 x 1 x 2 x 2 x 667) or 32016 MBps or 32 GBps. These values are maximum theoretical throughputs for comparison purposes only.

Table 2-3 provides the theoretical throughput values of 4.7 GHz processors and 667 Mhz memory configuration.

Table 2-3 Theoretical throughput values

Memory	Bandwidth
L1 (Data)	75.2 GB/sec
L2 / Chip	300.8 GB/sec
L3 / Chip	37.6 GB/sec
Memory / Chip	32 GB/sec
Inter-Node Buses (16-cores)	75.2 GB/sec
Intra-Node Buses (16-cores)	100.26 GB/sec

2.5 System buses

The following sections provide additional information related to the internal buses.

2.5.1 I/O buses and GX+ card

Each POWER6 processor provides a GX+ bus which is used to connect to an I/O subsystem or Fabric Interface card. The processor card populating the first processor slot is connected to the GX+ multifunctional host bridge chip which provides the following major interfaces:

- One GX+ passthru bus:

  GX+ passthru elastic interface runs at one half the frequency of the primary. It allows other GX+ bus hubs to be connected into the system.

- Two 64-bit PCI-X 2.0 buses, one 64-bit PCI-X 1.0 bus, and one 32-bit PCI-X 1.0 bus
- Four 8x PCI Express links
- Two 10 Gbps Ethernet ports: Each port is individually configurable to function as two 1 Gb/s port

In a fully populated 570, there are two GX+ buses, one from each processor. Each 570 has 2 GX+ slots with a single GX+ bus. The GX+ multifunctional host bridge provide a dedicated GX+ bus routed to the first GX+ slot through GX+ passthru bus. The second GX+ slot is not active unless the second processor card is installed. It is not required for CoD cards to be activated in order for the associated GX+ bus to be active.

Optional Dual port RIO-2 I/O Hub (FC 1800) and Dual port 12x Channel Attach (FC 1802) adapters that are installed in the GX+ slots are used for external DASD and IO drawer expansion. All GX+ cards are Hot-Pluggable.

Table 2-4 provides I/O bandwidth of 4.7 GHz processors configuration.

Table 2-4 I/O bandwidth

I/O	Bandwidth
Total I/O	62.6 GB/sec (16-cores)
Primary GX Bus	9.4 GB/sec (per node)
GX Bus Slot 1	4.7 GB/sec (per node)
GX Bus Slot 2	6.266 GB/sec (per node)

2.5.2 Service processor bus

The Service Processor (SP) flex cable is at the rear of the system and is used for SP communication between the system drawers. The SP cable remains similar to the p5 570 in that there is a unique SP cable for each configuration, but the p5 570 SP cables cannot be used for 570. Although, SP function is implemented in system drawer 1 and system drawer 2, Service interface card is required in every system drawer for signal distribution functions inside the system drawer. There is a unique SP cable for each drawer as Figure 2-9 on page 37 shows.

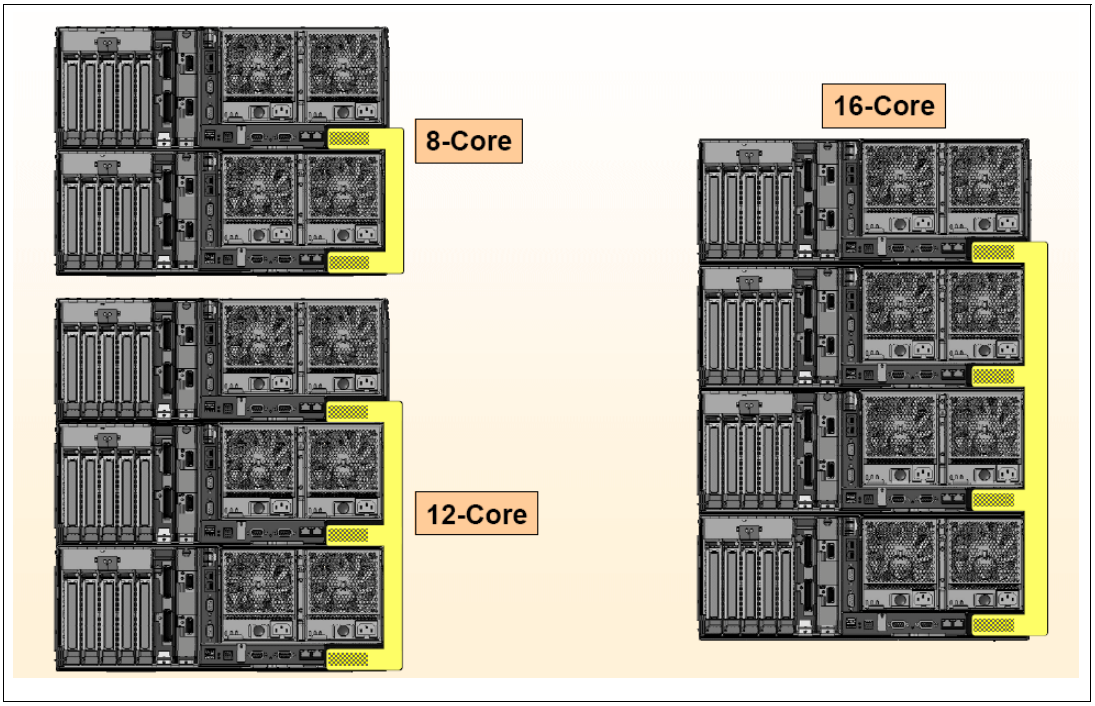


Figure 2-9 SP Flex™ cables

- FC 5657: 2-drawer SP cable
- FC 5658: 3-drawer SP cable
- FC 5660: 4-drawer SP cable.

2.6 Internal I/O subsystem

The internal I/O subsystem resides on the system planar which supports a mixture of both PCIe and PCI-X slots. All PCIe or PCI-X slots are hot pluggable and Enhanced Error Handling (EEH) enabled. In the unlikely event of a problem, EEH-enabled adapters respond to a special data packet generated from the affected PCIe or PCI-X slot hardware by calling system firmware, which will examine the affected bus, allow the device driver to reset it, and continue without a system reboot.

Table 2-5 display slot configuration of 570.

Table 2-5 Slot configuration of a 570

Slot#	Description	Location code	PHB	Max Card Size
Slot 1	PCIe x8	P1-C1	PCIe PHB0	Long
Slot 2	PCIe x8	P1-C2	PCIe PHB1	Long
Slot 3	PCIe x8	P1-C3	PCIe PHB2	Long
Slot 4	PCI-X DDR, 64-bit, 266 MHz	P1-C4	PCI-X PHB1	Long
Slot 5	PCI-X DDR, 64-bit, 266 MHz	P1-C5	PCI-X PHB3	Long

Slot#	Description	Location code	PHB	Max Card Size
Slot 6	PCIe x8 GX+	P1-C6 P1-C8	PCIe PHB3	Short
Slot 7	GX+	P1-C9		

Adapter slots P1-C6 and P1-C8 share the same physical space in a system enclosure. When a GX+ adapter is installed in GX slot P1-C8, PCIe slot P1-C6 cannot be used.

The 570 uses generation 3, blind swap cassettes to manage the installation and removal of adapters. Cassettes can be installed and removed without removing the drawer from the rack.

2.6.1 System ports

Although each system drawer is equipped with an Integrated Virtual Ethernet adapter (IVE) that has up to two serial ports, only the serial ports which are located on the system drawer (non-IVE) communicate with the service processor. They are called system ports. In operating system environment, the system ports become host virtual system ports and are not general RS232 serial port, but rather are limited use ports available for specifically supported functions.

The use of the integrated system ports on a 570 is limited to serial connected TTY console functionality and IBM approved call-home modems. These system ports do not support other general serial connection uses, such as UPS, HACMP heartbeat, printers, mice, track balls, space balls, etc. If you need serial port function, optional PCI adapters which are described in 2.8.6, 2.8.6, “Asynchronous PCI adapters” on page 47 are available.

If an HMC is connected, a *virtual serial console* is provided by the HMC (logical device vsa0 under AIX), and you can also connect a modem to the HMC. The system ports are not usable in this case. Either the HMC ports or the system ports can be used, but not both.

Configuration of the system ports, including basic ports settings (baud rate, etc.), modem selection, and call-home, can be accomplished with the Advanced Systems Management Interface (ASMI).

Note: The 570 must have an HMC. In normal operation, the system ports are for service representatives only.

2.7 Integrated Virtual Ethernet adapter

The POWER6 processor-based servers extend the virtualization technologies introduced in POWER5 by offering the Integrated Virtual Ethernet adapter (IVE). IVE, also named Host Ethernet Adapter (HEA) in other documentation, enables an easy way to manage the sharing of the integrated high-speed Ethernet adapter ports. It is a standard set of features that are part of every POWER6 processor-based server.

Integrated Virtual Ethernet adapter is a 570 standard feature but you can select from different options. The IVE comes from a general market requirements for improved performance and virtualization for Ethernet. It offers:

- Either two 10 Gbps Ethernet ports or four 1 Gbps ports or two 1 Gbps integrated ports
- A low cost Ethernet solution for low-end and mid-range System p servers

- Virtual Ethernet resources without the Virtual I/O Server
- Designed to operate at media speeds

The IVE is a physical Ethernet adapter that is connected directly to the GX+ bus instead of being connected to a PCIe or PCI-X bus, either as an optional or integrated PCI adapter. This provides IVE high throughput, and low latency. IVE also includes special hardware features to provide logical Ethernet adapters that can communicate to logical partitions (LPAR) reducing the use of POWER Hypervisor™ (PHYP).

IVE design provides a direct connection for multiple LPARs to share its resources. This allows LPARs to access external networks through the IVE without having to go through an Ethernet bridge on another logical partition, such as a Virtual I/O Server. Therefore, this eliminates the need to move packets (using virtual Ethernet) between partitions and then through a shared Ethernet adapter (SEA) to an Ethernet port. LPARs can share IVE ports with improved performance.

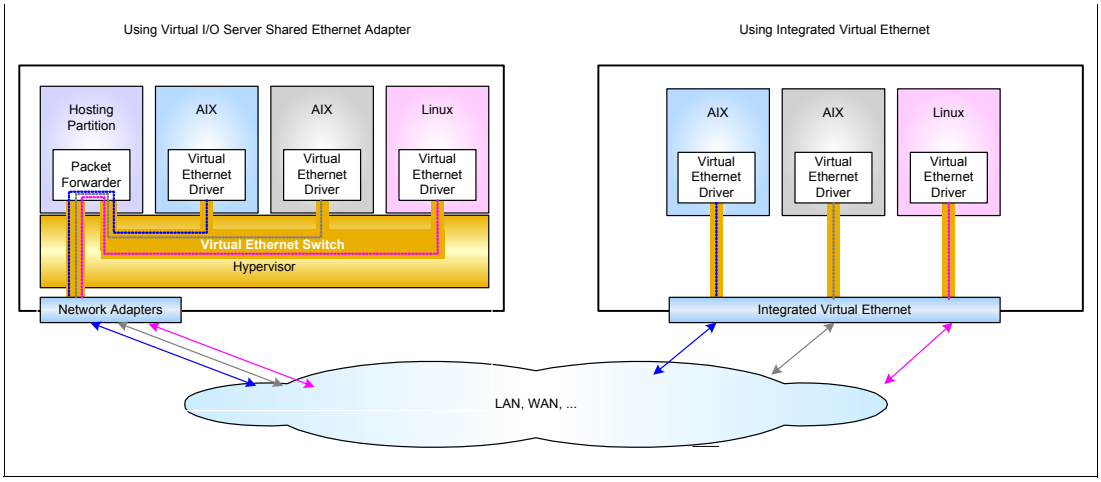


Figure 2-10 Integrated Virtual Ethernet compared to Virtual I/O Server Shared Ethernet Adapter

IVE supports 2 or 4 Ethernet ports running at 1 Gbps and 2 ports running at 10 Gbps depending on the IVE feature ordered. In the case of a 570 server, clients that are using 1 Gbps connection bandwidth in their IT infrastructure could move up to 10 Gbps infrastructure by adding a new 570 enclosure with the Integrated 2-ports 10 Gbps virtual Ethernet.

After any IBM System p initial order, you must use the MES[1] process to make any changes in the system configuration.

For more information on IVE features read *Integrated Virtual Ethernet Technical Overview and Introduction,* REDP-4340.

2.7.1 Physical ports and system integration

The following sections discuss the physical ports and the features available at the time of writing on a 570.

Each 570 enclosure can have unique Integrated Virtual Ethernet adapters, so a fully configured 570 server can be comprised of several different IVE feature codes.

[1] MES stands for Miscellaneous Equipment Shipment. It is the IBM process for IBM system upgrade

The following feature codes are available, at the time of writing, for each 570 enclosure:

- FC 5636 (standard), Integrated 2-ports 1 Gbps (single controller, twisted pair)
  - 16 MAC addresses, one port group
- FC 5637 (optional), Integrated 2-ports 10 Gbps SR[2] (single controller, optical)
  - 32 MAC addresses, two port groups
- FC 5639 (optional), Integrated 4-ports 1 Gbps (single controller, twisted pair)
  - 32 MAC addresses, two port groups

Figure 2-11 shows the major components of the Integrated Virtual Ethernet adapter hardware and additional system ports, according to the different feature codes.

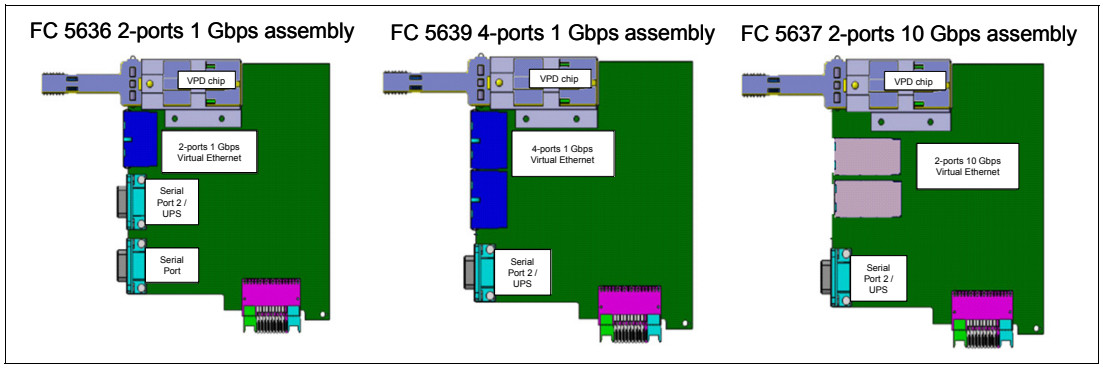

FC 5636 2-ports 1 Gbps assembly
VPD chip
2-ports 1 Gbps Virtual Ethernet
Serial Port 2 / UPS
Serial Port
FC 5639 4-ports 1 Gbps assembly
VPD chip
4-ports 1 Gbps Virtual Ethernet
Serial Port 2 / UPS
FC 5637 2-ports 10 Gbps assembly
VPD chip
2-ports 10 Gbps Virtual Ethernet
Serial Port 2 / UPS

Figure 2-11 Integrated Virtual Ethernet feature codes and assemblies

Any IVE feature code located in the first enclosure of a 570 also includes the System VPD (Vital Product Data) Chip and system (serial) port (1 or 2 depending on the feature code).

The IVE feature code is installed by manufacturing. Similar to other integrated ports, the feature does not have hot-swappable or hot-pluggable capability and must be serviced by a trained IBM System Service Representative.

Figure 2-12 on page 41 shows the rear view of a basic 570 in a state of disassembly with some necessary components and covers removed to highlight the connection of the feature code assembly into the server enclosure I/O subsystem system board.

[2] 10 Gbps SR (short range) is designed to support short distances over deployed multi-mode fiber cabling, it has a range of between 26 m and 82 m depending on cable type. It also supports 300 m operation over new, 50 µm 2000 MHz·km multi-mode fiber (using 850 nm).

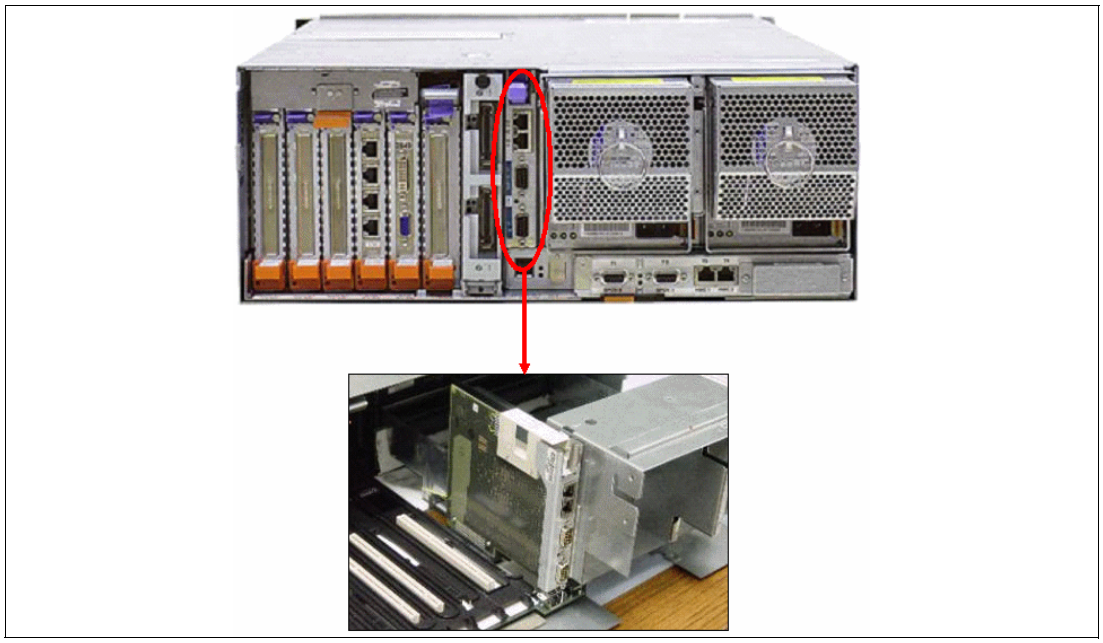


Figure 2-12 Integrated Virtual Ethernet adapter connection on System p 570 I/O system board

2.7.2 Feature code port and cable support

All the IVE feature codes have different connectivity options and different cable support (see Figure 2-13).

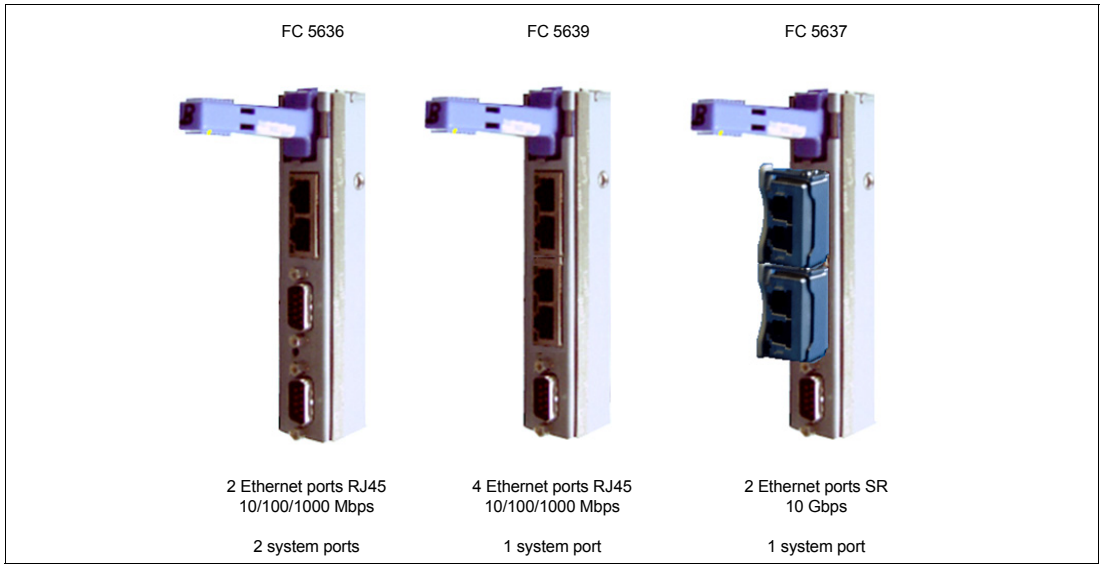


Figure 2-13 IVE physical port connectors according to IVE feature codes

FC 5636 and FC 5639 supports:

- 1 Gbps connectivity
- 10 Mbps and 100 Mbps connectivity
- RJ-45 connector

Use the Ethernet cables that meet Cat 5e[3] cabling standards, or higher, for best performance.

FC 5637 supports:

- Only 10 Gbps SR connectivity
- 62.5 micron multi-mode fiber cable type
  - LC physical connector type
  - 33 meters maximum range

2.7.3 IVE subsystem

Figure 2-14 shows a high level-logical diagram of the IVE.

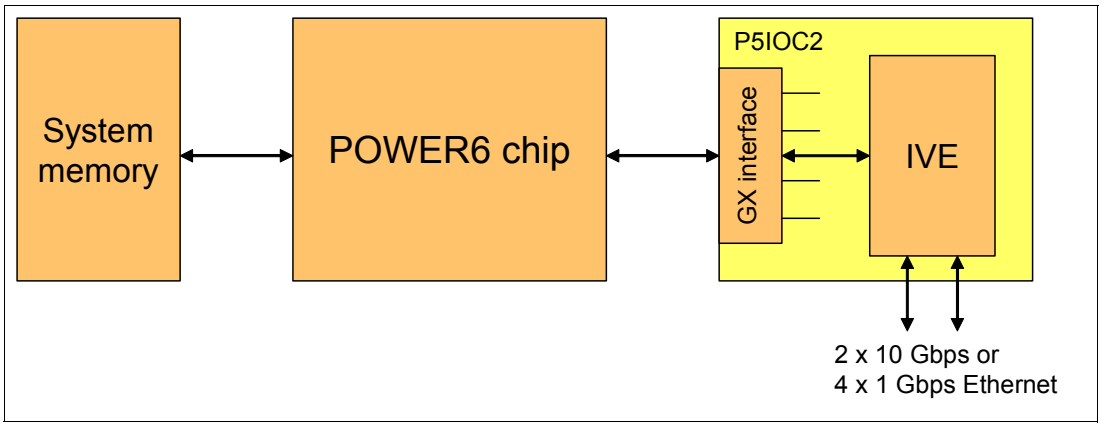


Figure 2-14 IVE system placement

Every POWER6 processor-based server I/O subsystem contains the P5IOC2 chip. It is a dedicated controller that acts as the primary bridge for all PCI buses and all internal I/O devices. IVE major hardware components reside inside the P5IOC2 chip.

The IVE design provides a great improvement of latency for short packets. Messaging applications such as distributed databases require low latency communication for synchronization and short transactions. The methods used to achieve low latency include:

- GX+ bus attachment
- Immediate data in descriptors (reduce memory access)
- Direct user space per-connection queueing (OS bypass)
- Designed for up to 3-times throughput improvement over current 10 Gbps solutions
- Provide additional acceleration functions in order to reduce host code path length. These include header / data split to help with zero-copy stacks
- Provide I/O virtualization support so that all partitions of the system can natively take advantage of the above features
- Allows one 10 Gbps port to replace up to 10 dedicated PCI 1 Gbps adapters in a partitioned system

One of the key design goals of the IVE is the capability to integrate up to two 10 Gbps Ethernet ports or four 1 Gbps Ethernet ports into the P5IOC2 chip, with the effect of a low cost Ethernet solution for low-end and mid-range server platforms. Any 10 Gbps, 1 Gbps, 100 Mbps or 10 Mbps speeds share the same I/O pins and do not require additional hardware

[3] Category 5 cable, commonly known as Cat 5, is a twisted pair cable type designed for high signal integrity. Category 5 has been superseded by the Category 5e specification.

or feature on top of the IVE card assembly itself. Another key goal is the support of all the state-of-art NIC functionality provided by leading Ethernet NIC vendors.

IVE offers the following functions with respect to virtualization:

- Up to 32 logical ports identified by MAC address
- Sixteen MAC addresses are assigned to each IVE port group.
- Each logical port can be owned by a separate LPAR
- Direct data path to LPAR
- Function enablement per LPAR
- Default send and receive queues per LPAR
- Ethernet MIB and RMON counters per LPAR
- VLAN filtering per logical port (4096 VLANs * 32 Logical Ports)
- Internal layer 2 switch for LPAR to LPAR data traffic
- Multicast / Broadcast redirection to Multicast / Broadcast manager

IVE relies exclusively on the system memory and CEC processing cores to implement acceleration features. There is not a requirement of dedicated memory, thus reducing the cost of this solution and providing maximum flexibility. IVE Ethernet MACs and acceleration features consume less than 8 mm^2 of logic in CMOS 9SF technology.

IVE does not have flash memory for its open firmware but it is stored in the Service Processor flash and then passed to POWER Hypervisor (PHYP) control. Therefore flash code update is done by PHYP.

2.8 PCI adapters

Peripheral Component Interconnect Express PCIe uses a serial interface and allows for point-to-point interconnections between devices using directly wired interface between these connection points. A single PCIe serial link is a dual-simplex connection using two pairs of wires, one pair for transmit and one pair for receive, and can only transmit one bit per cycle. It can transmit at the extremely high speed of 2.5 Gbps, which equates to a burst mode of 320 MBps on a single connection. These two pairs of wires is called a lane. A PCIe link may be comprised of multiple lanes. In such configurations, the connection is labeled as x1, x2, x8, x12, x16 or x32, where the number is effectively the number of lanes.

IBM offers PCIe adapter options for the 570, as well as PCI and PCI-extended (PCI-X) adapters. All adapters support Extended Error Handling (EEH). PCIe adapters use a different type of slot than PCI and PCI-X adapters. If you attempt to force an adapter into the wrong type of slot, you may damage the adapter or the slot. A PCI adapter can be installed in a PCI-X slot, and a PCI-X adapter can be installed in a PCI adapter slot. A PCIe adapter cannot be installed in a PCI or PCI-X adapter slot, and a PCI or PCI-X adapter cannot be installed in a PCIe slot. For a full list of the adapters that are supported on the systems and for important information regarding adapter placement, see *PCI Adapter Placement*, SA76-0090. You can find this publication at :

https://www-01.ibm.com/servers/resourcelink/lib03030.nsf/pages/pHardwarePublicationsByCategory?OpenDocument&pathID=

Before adding or rearranging adapters, use the System Planning Toll to validate the new adapter configuration. See the System Planning Tool Web site at :

http://www-03.ibm.com/servers/eserver/support/tools/systemplanningtool/

If you are installing a new feature, ensure that you have the software required to support the new feature and determine whether there are any existing PTF prerequisites to install. To do this, use the IBM Prerequisite Web site at :

http://www-912.ibm.com/e_dir/eServerPrereq.nsf

2.8.1 LAN adapters

To connect a 570 local area network (LAN), you can use Integrated 10/100/1000 dual-port Virtual Ethernet with optional 10/100/1000 quad-port or dual-port 10 Gb Virtual Ethernet. Table 2-6 lists the additional LAN adapters that are available.

Table 2-6 Available LAN adapters

Feature code	Adapter description	Slot	Size
5700	Gigabit Ethernet-SX	PCI-X	Short
5701	10/100/1000 Base-TX Ethernet	PCI-X	Short
5706	2-port 10/100/1000 Base-TX	PCI-X	Short
5707	2-port Gigabit Ethernet	PCI-X	Short
5717	4-port 1 Gb Ethernet PCIe 4x	PCIe	Short
5718[a]	10 Gigabit Ethernet-SR	PCI-X	Short
5719[a]	IBM 10 Gigabit Ethernet-SR	PCI-X	Short
5721	10 Gb Ethernet - Short Reach	PCI-X	Short
5722	10 GB Ethernet - Long Reach	PCI-X	Short
5740	4-port 10/100/1000 Ethernet	PCI-X	Short
5767	2-port 1Gb Ethernet (UTP)	PCIe	Short
5768	2-port 1Gb Ethernet (Fiber)	PCIe	Short

a. Supported, but not available for a new configuration

2.8.2 SCSI and SAS adapters

To connect to external SCSI or SAS devices, the adapters that are provided in Table 2-7 are available to be configured.

Table 2-7 Available SCSI adapters

Feature code	Adapter description	Slot	Size
5712[a]	Dual Channel Ultra320 SCSI	PCI-X	Short
5736	DDR Dual Channel Ultra320 SCSI	PCI-X	Short
5900	PCI-X DDR Dual -x4 SAS Adapter	PCI-X	Short

a. Supported, but not available for a new configuration

Table 2-8 on page 45 shows comparing Parallel SCSI to SAS.

Table 2-8 Comparing Parallel SCSI to SAS

	Parallel SCSI	SAS
Architecture	Parallel, all devices connected to shared bus	Serial, point-to-point, discrete signal paths
Performance	320 Mb/s (Ultra320 SCSI), performance degradeds as devices added to shared bus	3 Gb/s, roadmap to 12 Gb/s, performance maintained as more devices added
Scalability	15 drives	Over 16,000 drives
Compatibility	Incompatible with all other drive interfaces	Compatible with Serial ASA (SATA)
Max. Cable Length	12 meters total (must sum lengths of all cables used on bus)	8 meters per discrete connection, total domain cabling hundreds of meters
Cable From Factor	Multitude of conductors adds bulk, cost	Compact connectors and cabling save space, cost
Hot Pluggability	No	Yes
Device Identification	Manually set, user must ensure no ID number conflicts on bus	Worldwide unique ID set at time of manufacture
Termination	Manually set, user must ensure proper installation and functionality of terminators	Discrete signal paths enable device to include termination by default

2.8.3 iSCSI

iSCSI is an open, standards-based approach by which SCSI information is encapsulated using the TCP/IP protocol to allow its transport over IP networks. It allows transfer of data between storage and servers in block I/O formats (that is defined by iSCSI protocol) and thus enables the creation of IP SANs. iSCSI allows an existing network to transfer SCSI commands and data with full location independence and defines the rules and processes to accomplish the communication. The iSCSI protocol is defined in iSCSI IETF draft-20. For more information about this standard, see:

http://tools.ietf.org/html/rfc3720

Although iSCSI can be, by design, supported over any physical media that supports TCP/IP as a transport, today's implementations are only on Gigabit Ethernet. At the physical and link level layers, iSCSI supports Gigabit Ethernet and its frames so that systems supporting iSCSI can be directly connected to standard Gigabit Ethernet switches and IP routers. iSCSI also enables the access to block-level storage that resides on Fibre Channel SANs over an IP network using iSCSI-to-Fibre Channel gateways such as storage routers and switches. The iSCSI protocol is implemented on top of the physical and data-link layers and presents to the operating system a standard SCSI Access Method command set. It supports SCSI-3 commands and reliable delivery over IP networks. The iSCSI protocol runs on the host initiator and the receiving target device. It can either be optimized in hardware for better performance on an iSCSI host bus adapter (such as FC 5713 and FC 5714 supported in IBM System p servers) or run in software over a standard Gigabit Ethernet network interface card. IBM System p systems support iSCSI in the following two modes:

Hardware Using iSCSI adapters (see “IBM iSCSI adapters” on page 46).

Software Supported on standard Gigabit adapters, additional software (see “IBM iSCSI software Host Support Kit” on page 46) must be installed. The main processor is utilized for processing related to the iSCSI protocol.

Initial iSCSI implementations are targeted at small to medium-sized businesses and departments or branch offices of larger enterprises that have not deployed Fibre Channel SANs. iSCSI is an affordable way to create IP SANs from a number of local or remote storage devices. If Fibre Channel is present, which is typical in a data center, it can be accessed by the iSCSI SANs (and vice versa) via iSCSI-to-Fibre Channel storage routers and switches.

iSCSI solutions always involve the following software and hardware components:

Initiators These are the device drivers and adapters that reside on the client. They encapsulate SCSI commands and route them over the IP network to the target device.

Targets The target software receives the encapsulated SCSI commands over the IP network. The software can also provide configuration support and storage-management support. The underlying target hardware can be a storage appliance that contains embedded storage, and it can also be a gateway or bridge product that contains no internal storage of its own.

IBM iSCSI adapters

iSCSI adapters in IBM System p systems provide the advantage of increased bandwidth through the hardware support of the iSCSI protocol. The 1 Gigabit iSCSI TOE (TCP/IP Offload Engine) PCI-X adapters support hardware encapsulation of SCSI commands and data into TCP and transports them over the Ethernet using IP packets. The adapter operates as an iSCSI TOE. This offload function eliminates host protocol processing and reduces CPU interrupts. The adapter uses a Small form factor LC type fiber optic connector or a copper RJ45 connector.

Table 2-9 provides the orderable iSCSI adapters.

Table 2-9 Available iSCSI adapters

Feature code	Description	Slot	Size
5713	Gigabit iSCSI TOE on PCI-X on copper media adapter	PCI-X	Short
5714	Gigabit iSCSI TOE on PCI-X on optical media adapter	PCI-X	Short

IBM iSCSI software Host Support Kit

The iSCSI protocol can also be used over standard Gigabit Ethernet adapters. To utilize this approach, download the appropriate iSCSI Host Support Kit for your operating system from the IBM NAS support Web site at:

`http://www.ibm.com/storage/support/nas/`

The iSCSI Host Support Kit on AIX and Linux acts as a software iSCSI initiator and allows you to access iSCSI target storage devices using standard Gigabit Ethernet network adapters. To ensure the best performance, enable the TCP Large Send, TCP send and receive flow control, and Jumbo Frame features of the Gigabit Ethernet Adapter and the iSCSI Target. Tune network options and interface parameters for maximum iSCSI I/O throughput on the operating system.

2.8.4 Fibre Channel adapter

The 570 servers support direct or SAN connection to devices using Fibre Channel adapters. Table 2-10 provides a summary of the available Fibre Channel adapters.

All of these adapters have LC connectors. If you are attaching a device or switch with an SC type fibre connector, then an LC-SC 50 Micron Fiber Converter Cable (FC 2456) or an LC-SC 62.5 Micron Fiber Converter Cable (FC 2459) is required.

Supported data rates between the server and the attached device or switch are as follows: Distances of up to 500 meters running at 1 Gbps, distances up to 300 meters running at 2 Gbps data rate, and distances up to 150 meters running at 4 Gbps. When these adapters are used with IBM supported Fibre Channel storage switches supporting long-wave optics, distances of up to 10 kilometers are capable running at 1 Gbps, 2 Gbps, and 4 Gbps data rates.

Table 2-10 Available Fibre Channel adapters

Feature code	Description	Slot	Size
5716[a]	2 Gigabit Fibre Channel PCI-X Adapter	PCI-X	Short
5758	DDR 4 Gb single port Fibre Channel	PCI-X	Short
5759	DDR 4 Gb dual port Fibre Channel	PCI-X	Short
5773	1-port 4 Gb Fibre Channel	PCIe	Short
5774	2-port 4 Gb Fibre Channel	PCIe	Short

a. Supported, but not available for a new configuration

2.8.5 Graphic accelerators

The 570 support up to four graphics adapters. Table 2-11 provides the available graphic accelerators. They can be configured to operate in either 8-bit or 24-bit color modes. These adapters support both analog and digital monitors.

Table 2-11 Available Graphic accelerators

Feature code	Description	Slot	Size
2849	GXT135P Graphics Accelerator	PCI-X	Short
5748	GXT145 Graphics Accelerator	PCIe	Short

Note: Both adapters are not hot-pluggable.

2.8.6 Asynchronous PCI adapters

Asynchronous PCI-X adapters provide connection of asynchronous EIA-232 or RS-422 devices. If you have a cluster configuration or high-availability configuration and plan to connect the IBM System p servers using a serial connection, the use of the two system ports is not supported. You should use one of the features listed in Table 2-12 on page 48.

Table 2-12 Asynchronous PCI-X adapters

Feature code	Description	Slot	Size
2943	8-Port Asynchronous Adapter EIA-232/RS-422	PCI-X	Short
5723	2-Port Asynchronous IEA-232 PCI Adapter	PCI-X	Short

In many cases, the FC 5723 asynchronous adapter is configured to supply a backup HACMP heartbeat. In these cases, a serial cable (FC 3927 or FC 3928) must be also configured. Both of these serial cables and the FC 5723 adapter have 9-pin connectors.

2.8.7 Additional support for existing PCI adapters

The lists of the major PCI adapters that you can configure in a 570 when you build an available configuration are described in 2.8.1, “LAN adapters” on page 44 through 2.8.6, “Asynchronous PCI adapters” on page 47. The list of all the supported PCI adapters, with the related support for additional external devices, is more extended.

If you would like to use PCI adapters you already own, contact your IBM service representative to verify whether those adapters are supported.

2.9 Internal storage

The 570 internal disk subsystem is driven by the latest DASD interface technology Serial Attached SCSI (SAS). This interface provides enhancements over parallel SCSI with its point to point high frequency connections. The SAS controller has eight SAS ports, four of them are used to connect to the DASD drives and one to a media device.

The DASD backplane implements two SAS port expanders that take four SAS ports from the SAS controller and expands it to 12 SAS ports. These 12 ports allow for redundant SAS ports to each of the six DASD devices.

The DASD backplane provides the following functions :

- supports six 3.5 inches SAS DASD devices
- contains two SAS port expanders for redundant SAS paths to the SAS devices
- SAS passthru connection to medias backplane

2.9.1 Integrated RAID options

The 570 supports a 6-pack DASD backplane attached to the system planar. To support RAID functionality a combination of additional adapters is required. At the time of writing RAID level 0 and 10 are supported using adapter FC 5900 or 5909 and FC 3650 or FC 3651 as described in Table 2-13.

Table 2-13 Raid support

Feature	Description
3650 + 5900 or 5909	A hardware feature that occupies PCIe slot P1C3 provides a mini SAS 4x connector to the rear bulkhead and allows three of the internal SAS drives (disk 4,5,6) to be controlled using an external SAS controller.

Feature	Description
3651 + 5900 or 5909	A hardware feature that occupies PCIe slot P1C3 provides a mini SAS 4x connector to the rear bulkhead and allows all of the internal SAS drives to be controlled using an external SAS controller.

2.9.2 Split backplane

As described in paragraph 2.9.1, “Integrated RAID options” on page 48 the same features are required to split the 6-pack DASD backplane in two groups of three disks. Using feature 3650 and 5900 disk 4,5, and, 6 are managed using the external SAS controller. Disk 1, 2, and 3 are managed by the internal SAS controller.

2.9.3 Internal media devices

Inside each CEC drawer in the 570 there is an optional media backplane with one media bay. The internal IDE media bay in separate CEC drawers can be allocated or assigned to a different partition. The media backplane inside each CEC drawer cannot be split between two logical partitions.

2.9.4 Internal hot-swappable SAS drives

The 570 can have up to six hot-swappable disk drives plugged in the physical 6-pack disk drive backplane. The hot-swap process is controlled by the virtual SAS Enclosure Services (VSES), which is located in the logical 6-pack disk drive backplane. The 6-pack disk drive backplanes can accommodate the devices listed in Table 2-14

Table 2-14 Hot-swappable disk options

Feature code	Description
3646	73.4 GB 15,000 RPM SAS hot-swappable disk drive
3647	146.8 GB 15,000 RPM SAS hot-swappable disk drive
3548	300 GB 15,000 RPM SAS hot-swappable disk drive

Prior to the hot-swap of a disk drive in the hot-swappable-capable bay, all necessary operating system actions must be undertaken to ensure that the disk is capable of being deconfigured. After the disk drive has been deconfigured, the SAS enclosure device will power-off the slot, enabling safe removal of the disk. You should ensure that the appropriate planning has been given to any operating-system-related disk layout, such as the AIX Logical Volume Manager, when using disk hot-swap capabilities. For more information, see *Problem Solving and Troubleshooting in AIX 5L*, SG24-5496.

Note: We recommend that you follow this procedure, after the disk has been deconfigured, when removing a hot-swappable disk drive:

1. Release the tray handle on the disk assembly.
2. Pull out the disk assembly a little bit from the original position.
3. Wait up to 20 seconds until the internal disk stops spinning.

Now you can safely remove the disk from the DASD backplane.

2.10 External I/O subsystems

This section describes the external I/O subsystems, which include the I/O drawers, the IBM System Storage™ EXP 12S SAS drawer, as well as the 7311-D11, 7311-D20, 7311-G30, 7031-D24, and 7031-T24 deskside tower.

Table 2-15 provided an overview of all the supported I/O drawers.

Table 2-15 I/O drawer capabilities

Drawer	DASD	PCI Slots	Requirements for a 570
7311-D11		6 x PCI-X	GX+ adapter card FC 1800
7311-D20	12 x SCSI disk drive bays	7 x PCI-X	GX+ adapter card FC 1800
7314-G30		6 x PCI-X DDR 266 MHz	GX+ adapter card FC 1802
7031-T24/D24	24 x SCSI disk drive bays		Any supported SCSI adapter
FC 5886	12 x SAS disk drive bays		Any supported SAS adapter

Each POWER6 chip provides a GX+ bus which is used to connect to an I/O subsystem or Fabric Interface card. In a fully populated 570 enclosure there are two GX+ buses, one from each POWER6 chip. Each 570 enclosure has 2 GX+ slots with a single GX+ bus. The second GX+ slot is not active unless the second CPU card is installed. If the second CPU card is installed, then the second GX+ slot and associated bus is active and available.

The maximum number of attached remote I/O drawers depends on the number of system unit enclosures in the system and the I/O attachment type. Each GX+ bus can be populated with a GX+ adapter card that adds more RIO-G ports to connect external I/O drawers.

The GX+ adapter cards listed in Table 2-16 are supported at the time of writing.

Table 2-16 GX+ adapter card supported

Feature code	GX+ adapter card description	GX+ adapter card I/O drawer support
FC 1800	GX dual port RIO-2 attach	Card provides two ports that support up to four of the following I/O drawers: ► 7311-D10 ► 7311-D11 ► 7311-D20
FC 1802	GX dual port 12X channel attach	Card provides two 12X connections that support up to four of the following I/O drawer: ► 7314-G30

2.10.1 7311 Model D11 I/O drawers

The 7311-D11 provides six PCI-X slots supporting an enhanced blind-swap mechanism. Drawers must have a RIO-2 adapter to connect to the server.

Each primary PCI-X bus is connected to a PCI-X-to-PCI-X bridge, which provides three slots with Extended Error Handling (EEH) for error recovering. In the 7311 Model D11 I/O drawer,

slots 1 to 6 are PCI-X slots that operate at 133 MHz and 3.3 V signaling. Figure 2-15 on page 51 shows a conceptual diagram of the 7311 Model D11 I/O drawer.

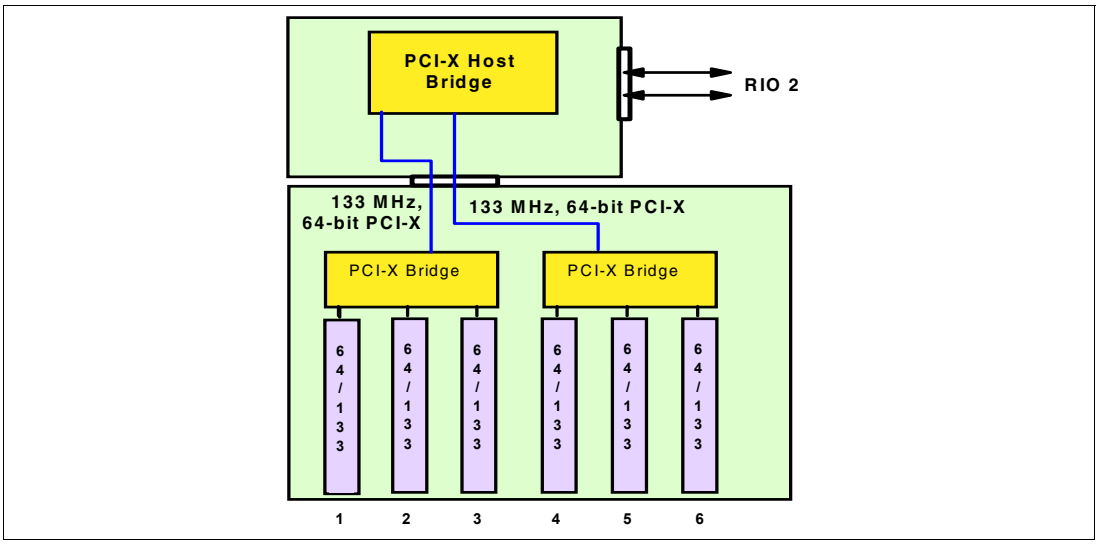


Figure 2-15 Conceptual diagram of the 7311-D11 I/O drawer

7311 Model D11 features

This I/O drawer model provides the following features:

- Six hot-plug 64-bit, 133 MHz, 3.3 V PCI-X slots, full length, enhanced blind-swap cassette
- Default redundant hot-plug power and cooling
- Two default remote (RIO-2) ports and two SPCN ports

7311 Model D11 rules and maximum support

Table 2-17 describes the maximum number of I/O drawer supported

Table 2-17 Maximum number of 7311 Model D11 I/O drawers supported

570 enclosures	CPU card quantity	Max GX+ adapter card	Max I/O drawer supported
1	1	1	4
1	2	2	8
2	4	4	12
3	6	6	16
4	8	8	20

2.10.2 Consideration for 7311 Model D10 I/O drawer

It is not possible to configure the 7311 Model D10 I/O drawer in a 570 initial order. Clients who decided to migrate p5 570 to 570 can decide to re-use 7311 Model D10 I/O drawers originally connected to the p5 570. It requires the same connection of the 7311 Model D11 and can be intermixed in the same loop

2.10.3 7311 Model D20 I/O drawer

The 7311 Model D20 I/O drawer must have the RIO-2 loop adapter (FC 6417) to be connected to the 570 system. The PCI-X host bridge inside the I/O drawer provides two primary 64-bit PCI-X buses running at 133 MHz. Therefore, a maximum bandwidth of 1 GBps is provided by each of the buses.

Figure 2-16 shows a conceptual diagram of the 7311 Model D20 I/O drawer subsystem.

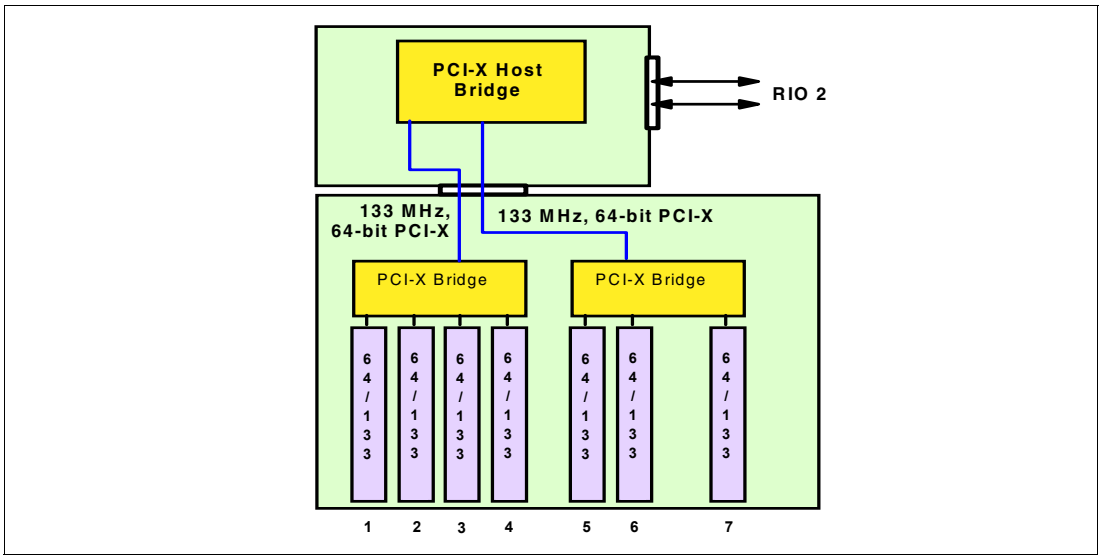


Figure 2-16 Conceptual diagram of the 7311-D20 I/O drawer

7311 Model D20 internal SCSI cabling

A 7311 Model D20 supports hot-swappable SCSI Ultra320 disk drives using two 6-pack disk bays for a total of 12 disks. Additionally, the SCSI cables (FC 4257) are used to connect a SCSI adapter (any of various features) in slot 7 to each of the 6-packs, or two SCSI adapters, one in slot 4 and one in slot 7. (See Figure 2-17.)

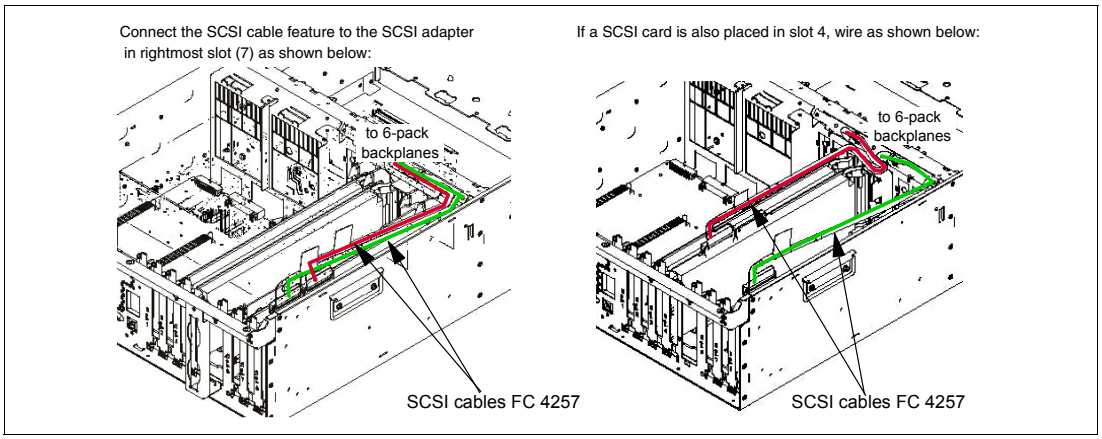

Connect the SCSI cable feature to the SCSI adapter in rightmost slot (7) as shown below:
to 6-pack backplanes
SCSI cables FC 4257
If a SCSI card is also placed in slot 4, wire as shown below:
to 6-pack backplanes
SCSI cables FC 4257

Figure 2-17 7311 Model D20 internal SCSI cabling

Note: Any 6-packs and the related SCSI adapter can be assigned to a logical partition (for example, the two partitions can be two Virtual I/O server partitions). If one SCSI adapter is connected to both 6-packs, then both 6-packs can be assigned only to the same partition.

2.10.4 7311 Model D11 and Model D20 I/O drawers and RIO-2 cabling

As described in 2.10, “External I/O subsystems” on page 50, we can connect up to four I/O drawers in the same loop, and up to 20 I/O drawers to the p5-570 system.

Each RIO-2 port can operate at 1 GHz in bidirectional mode and is capable of passing data in each direction on each cycle of the port. Therefore, the maximum data rate is 4 GBps per I/O drawer in double barrel mode (using two ports).

There is one default primary RIO-2 loop in any 570 building block. This feature provides two Remote I/O ports for attaching up to four 7311 Model D11 or 7311 Model D20 I/O drawers or 7311 Model D10 to the system in a single loop. Different I/O drawer models can be used in the same loop, but the combination of I/O drawers must be a total of four per single loop. The optional RIO-2 expansion card may be used to increase the number of I/O drawers that can be connected to one 570 building block, and the same rules of the default RIO-2 loop must be considered. The method that is used to connect the drawers to the RIO-2 loop is important for performance.

Figure 2-18 shows how you could connect four I/O drawers to one p5-570 building block. This is a logical view; actual cables should be wired according to the installation instructions.

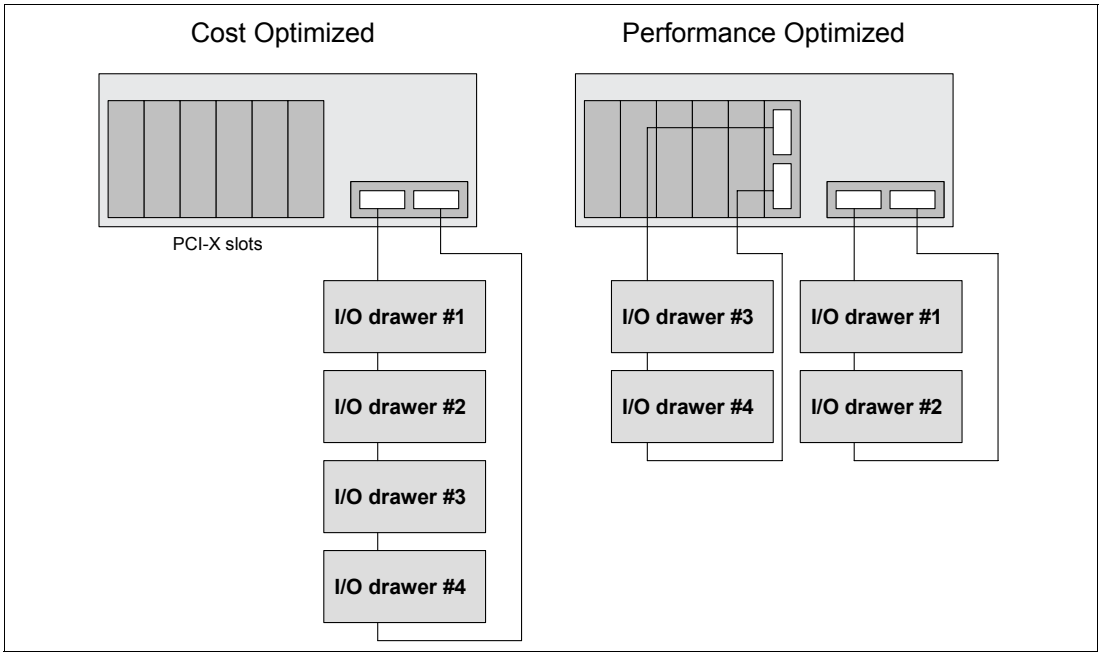


Figure 2-18 RIO-2 cabling examples

Note: If you have 20 I/O drawers, although there are no restrictions on their placement, this can affect performance.

RIO-2 cables are available in different lengths to satisfy different connection requirements:

- Remote I/O cable, 1.2 m (FC 3146, for between D11 drawers only)
- Remote I/O cable, 1.75 m (FC 3156)
- Remote I/O cable, 2.5 m (FC 3168)
- Remote I/O cable, 3.5 m (FC 3147)
- Remote I/O cable, 10 m (FC 3148)

2.10.5 7311 I/O drawer and SPCN cabling

SPCN[4] is used to control and monitor the status of power and cooling within the I/O drawer. SPCN is a loop: Cabling starts from SPCN port 0 on the 570 to SPCN port 0 on the first I/O drawer. The loop is closed, connecting the SPCN port 1 of the I/O drawer back to port 1 of the 570 system. If you have more than one I/O drawer, you continue the loop, connecting the next drawer (or drawers) with the same rule.

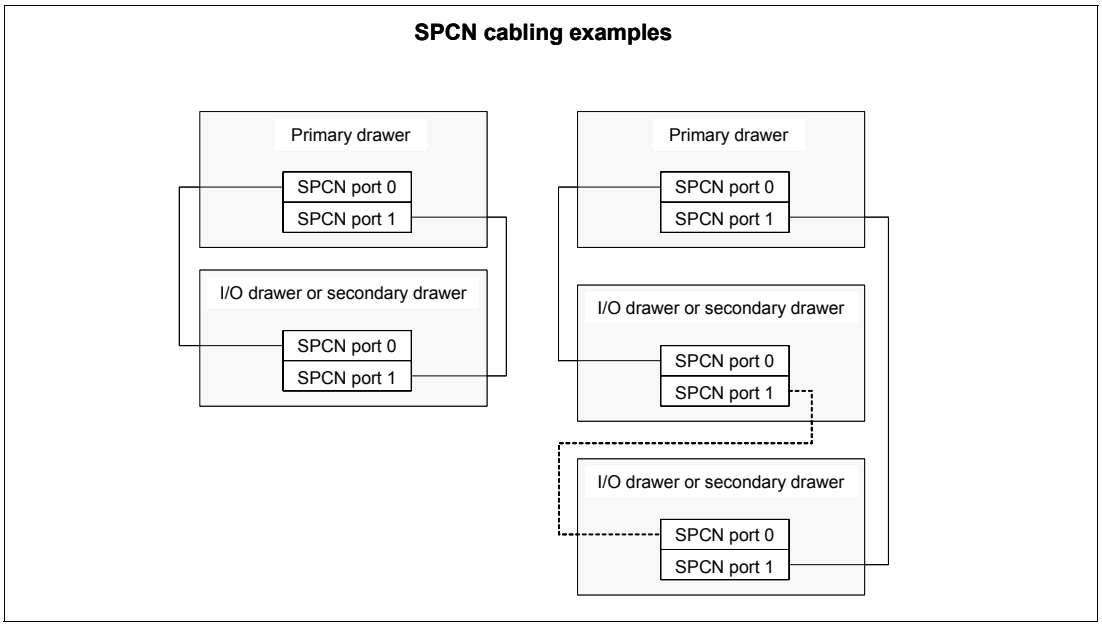


Figure 2-19 SPCN cabling examples

There are different SPCN cables to satisfy different length requirements:

- SPCN cable drawer-to-drawer, 2 m (FC 6001)
- SPCN cable drawer-to-drawer, 3 m (FC 6006)
- SPCN cable rack-to-rack, 6 m (FC 6008)
- SPCN cable rack-to-rack, 15 m (FC 6007)
- SPCN cable rack-to-rack, 30 m (FC 6029)

2.10.6 7314 Model G30 I/O drawer

The 7314-G30 expansion unit is a rack-mountable, I/O expansion drawer that is designed to be attached to the system unit using the InfiniBand® bus and InfiniBand cables. The 7314-G30 can accommodate 6 blind swap adapter cassettes. Cassettes can be installed and removed without removing the drawer from the rack. The Figure 2-20 on page 55 shows the back view of the expansion unit.

[4] System Power Control Network

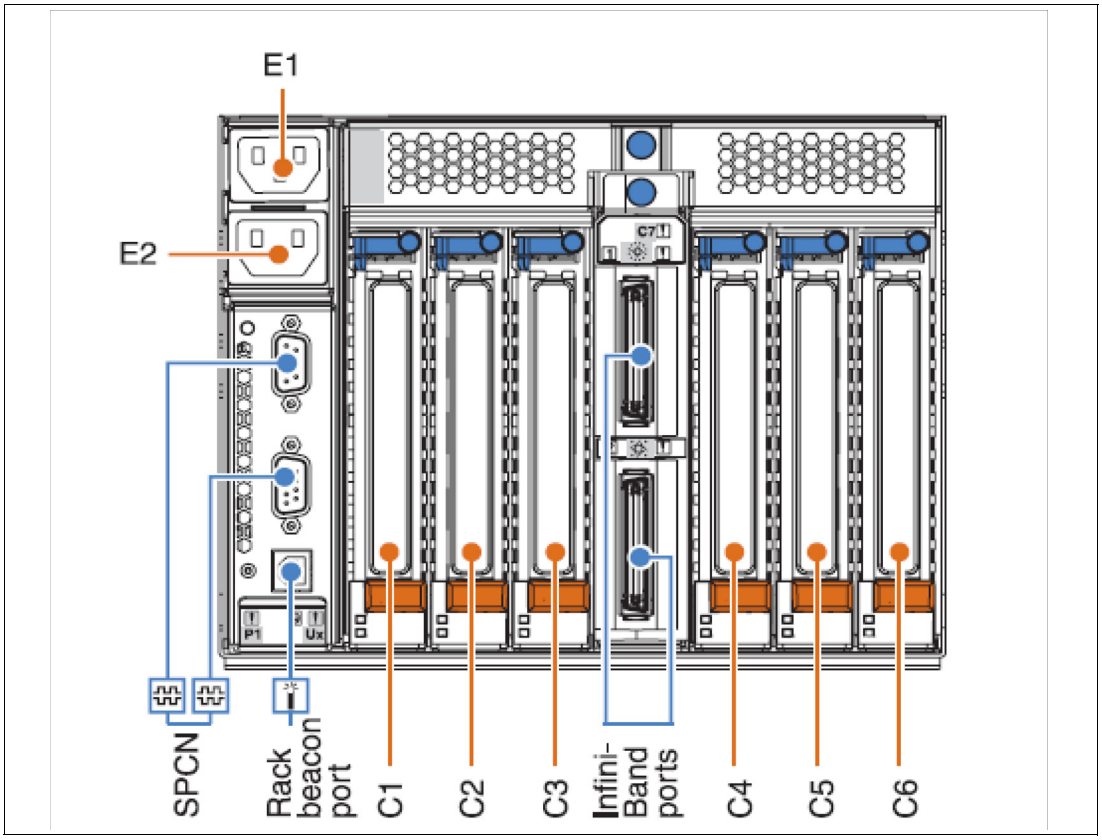


Figure 2-20 7311 Model G40 rear side

7314 Model G30 rules and maximum support

Table 2-18 describes the maximum number of I/O drawer supported

Table 2-18 Maximum number of 7314 Model G30 I/O drawers supported

570 enclosures	CPU card quantity	Max GX+ adapter card	Max I/O drawer supported
1	1	1	4
1	2	2	8
2	4	4	16
3	6	6	24
4	8	8	32

Similar to the 7311 Model D10 and D11, up to two 7314 Model G30 drawers can be installed in a unit enclosure (FC 7314). The unit enclosure requires to be installed in a 19" rack such as the 7014-T00 or 7014-T42. The actual installation location in the rack will vary depending on other rack content specify codes ordered with rack.

2.11 External disk subsystems

The 570 has internal hot-swappable drives. When the AIX operating system is installed in a IBM System p server, the internal disks are usually used for the AIX rootvg volume group and

paging space. Specific client requirements can be satisfied with the several external disk possibilities that the 570 supports.

2.11.1 IBM System Storage EXP 12S (FC 5886)

The IBM System Storage EXP 12S is a high density rack-mountable disk drive enclosure for supporting a total of twelve 3.5-inch disk drives on POWER6 systems only. Using hot-swappable 300 GB SAS disk drives the EXP 12S drawer can provide 3.6 TB of disk capacity. The expansion drawer provides redundant power device, cooling, and SAS expanders all devices are hot-swappable. The SAS disks are front accessible, using the same disk carrier and 3.5 inch SAS disk drives as used in IBM POWER6 Power Systems. As a two unit, 19-inch rack-mountable disk enclosure supports 73 GB, 146 GB, and 300 GB hot-swappable SAS disk drives.

The IBM System Storage EXP 12S drawer offers :

- Modular SAS disk expansion drawer
- Up to 12 3.5-inch SAS disk drives
- Variety of supported connection options, from single attachment to a HACMP solution
- Redundant hot-plug power and cooling with dual line cords
- Redundant and hot-swappable SAS expanders

IBM System Storage EXP 12S drawer physical description

The EXP 12S drawer must be mounted in a 19-inch rack, such as the IBM 7014-T00 or 7014-T42. The EXP 12S drawer has the following attributes:

- One drawer EXP 12S
  - Width: 481.76 mm (18.97 in)
  - Depth: 511.00 mm (20.12 in)
  - Height: 87.36 mm (3.38 in)
  - Weight: 18 kb (39.70 lp)

Connecting a EXP 12S drawer to a POWER6 system

To connect a EXP 12S SAS drawer to a POWER6 system an additional adapter is needed. Table 2-19 provides the current list of available adapters.

Table 2-19 SAS adapters

Feature	Description
5900	PCI-X DDR dual SAS adapter

Depending on the required configuration a different set of cables are needed to connect the EXP 12S drawer to the system or drawer. A list of cables are provided in Table 2-20.

Table 2-20 SAS connection cables

Feature	Description
3652	SAS cable (EE) drawer to drawer 1 meter
3652	SAS cable (EE) drawer to drawer 3 meter
3654	SAS cable (EE) drawer to drawer 3 meter

Feature	Description
3691	SAS cable (YO) adapter to SAS enclosure 1.5 meter
3692	SAS cable (YO) adapter to SAS enclosure 3 meter
3693	SAS cable (YO) adapter to SAS enclosure 6 meter

A typical base configuration is using a server machine and a single attached drawer as shown in Figure 2-21.

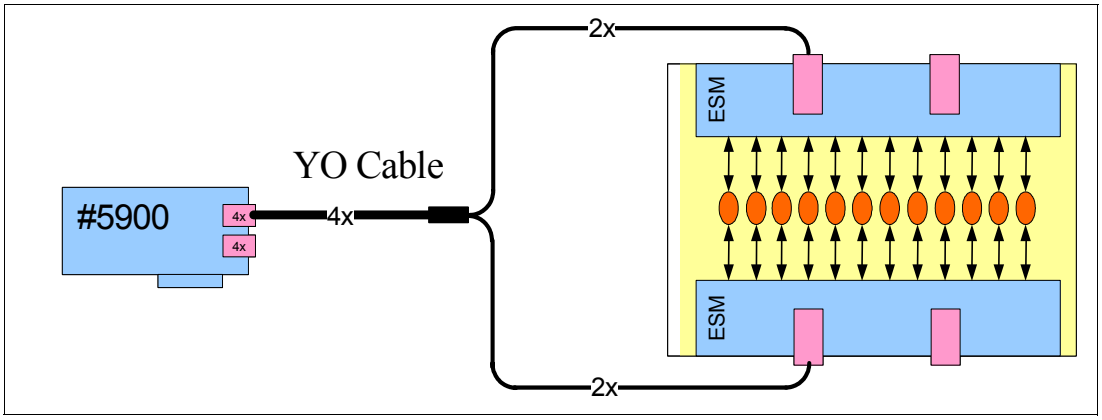


Figure 2-21 Base configuration of one SAS drawer

A maximum configuration using four EXP 12S drawers on one adapter feature is shown in Figure 2-22.

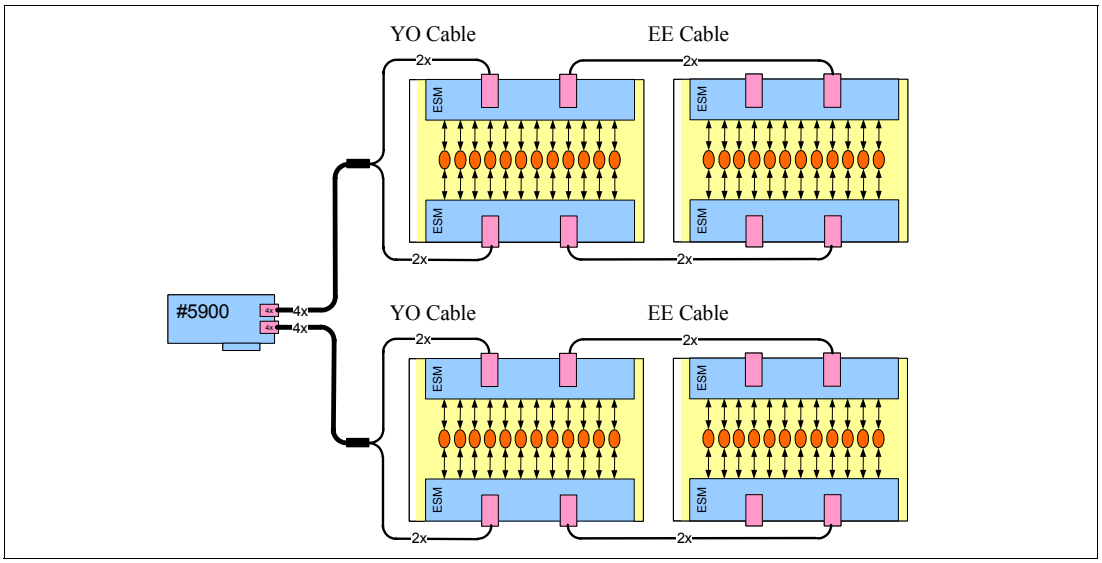


Figure 2-22 Maximum attachment of EXP 12S on one adapter

2.11.2 IBM TotalStorage EXP24 Expandable Storage

The IBM TotalStorage® EXP24 Expandable Storage disk enclosure, Model D24 or T24, can be purchased together with the 570 and will provide low-cost Ultra320 (LVD) SCSI disk storage. This disk storage enclosure device provides more than 7 TB of disk storage in a 4 U rack-mount (Model D24) or compact deskside (Model T24) unit. Whether high availability

storage solutions or simply high capacity storage for a single server installation, the unit provides a cost-effective solution. It provides 24 hot-swappable disk bays, 12 accessible from the front and 12 from the rear. Disk options that can be accommodated in any of the four six-packs disk drive enclosure are 73.4 GB, 146.8 GB, or 300 GB 10 K rpm or 36.4 GB, 73.4 GB, or 146.8 GB 15 K rpm drives. Each of the four six-packs disk drive enclosure might be attached independently to an Ultra320 SCSI or Ultra320 SCSI RAID adapter. For high available configurations, a dual bus repeater card (FC 5742) allows each six-pack to be attached to two SCSI adapters, installed in one or multiple servers or logical partitions. Optionally, the two front or two rear six-packs might be connected together to form a single Ultra320 SCSI bus of 12 drives.

2.11.3 IBM System Storage N3000, N5000 and N7000

The IBM System Storage N3000 and N5000 line of iSCSI enabled storage offerings provide a flexible way to implement a Storage Area Network over an Ethernet network. Flexible-Fibre Channel and SATA disk drive capabilities allow for deployment in multiple solution environments, including data compliant retention, nearline storage, disk-to-disk backup scenarios, and high-performance mission-critical I/O intensive operations.

Newest member of the IBM System storage N series family are the N7000 systems. The N7000 series is designed to deliver midrange to high-end enterprise storage and data management capabilities.

See the following link for more information:

http://www.ibm.com/servers/storage/nas

2.11.4 IBM TotalStorage Storage DS4000 Series

The IBM System Storage DS4000™ line of Fibre Channel enabled Storage offerings provides a wide range of storage solutions for your Storage Area Network. The IBM TotalStorage DS4000 Storage server family consists of the following models: DS4100, DS4300, DS4500, and DS4800. The Model DS4100 Express Model is the smallest model and scales up to 44.8 TB; the Model DS4800 is the largest and scales up to 89.6 TB of disk storage at the time of this writing. Model DS4300 provides up to 16 bootable partitions, or 64 bootable partitions if the turbo option is selected, that are attached with the Gigabit Fibre Channel Adapter (FC 1977). Model DS4500 provides up to 64 bootable partitions. Model DS4800 provides 4 GB switched interfaces. In most cases, both the IBM TotalStorage DS4000 family and the IBM System p5 servers are connected to a storage area network (SAN). If only space for the rootvg is needed, the Model DS4100 is a good solution.

For support of additional features and for further information about the IBM TotalStorage DS4000 Storage Server family, refer to the following Web site:

http://www.ibm.com/servers/storage/disk/ds4000/index.html

2.11.5 IBM System Storage DS6000 and DS8000 series

The IBM System Storage Models DS6000™ and DS8000™ are the high-end premier storage solution for use in storage area networks and use POWER technology-based design to provide fast and efficient serving of data. The IBM TotalStorage DS6000 provides enterprise class capabilities in a space-efficient modular package. It scales to 57.6 TB of physical storage capacity by adding storage expansion enclosures. The Model DS8000 series is the flagship of the IBM DS family. The DS8000 scales to 1024 TB. However, the system architecture is designed to scale to over one petabyte. The Model DS6000 and DS8000 systems can also be used to provide disk space for booting LPARs or partitions using

Micro-Partitioning technology. System Storage and the IBM Power servers are usually connected together to a storage area network.

For further information about ESS, refer to the following Web site:

http://www.ibm.com/servers/storage/disk/enterprise/ds_family.html

2.12 Hardware Management Console

The Hardware Management Console (HMC) is a dedicated workstation that provides a graphical user interface for configuring, operating, and performing basic system tasks for the POWER6 processor-based (as well as the POWER5 and POWER5+ processor-based) systems that function in either non-partitioned, partitioned, or clustered environments. In addition the HMC is used to configure and manage partitions. One HMC is capable of controlling multiple POWER5, POWER5+, and POWER6 processor-based systems.

At the time of writing, one HMC supports up to 48 POWER5, POWER5+ and POWER6 processor-based systems and up to 254 LPARs using the HMC machine code Version 7.3. For updates of the machine code and HMC functions and hardware prerequisites, refer to the following Web site:

https://www14.software.ibm.com/webapp/set2/sas/f/hmc/home.html

POWER5, POWER5+ and POWER6 processor-based system HMCs require Ethernet connectivity between the HMC and the server's service processor, moreover if dynamic LPAR operations are required, all AIX 5L, AIX V6, and Linux partitions must be enabled to communicate over the network to the HMC. Ensure that at least two Ethernet ports are available to enable public and private networks:

- The HMC 7042 Model C06 is a deskside model with one integrated 10/100/1000 Mbps Ethernet port, and two additional PCI slots.
- The HMC 7042 Model CR4 is a 1U, 19-inch rack-mountable drawer that has two native 10/100/1000 Mbps Ethernet ports and two additional PCI slots.

Note: The IBM 2-Port 10/100/1000 Base-TX Ethernet PCI-X Adapter (FC 5706) should be ordered to provide additional physical Ethernet connections.

For any logical partition in a server, it is possible to use a Shared Ethernet Adapter set in Virtual I/O Server or Logical Ports of the Integrated Virtual Ethernet card, for a unique or fewer connections from the HMC to partitions. Therefore, a partition does not require it's own physical adapter to communicate to an HMC.

It is a good practice to connect the HMC to the first HMC port on the server, which is labeled as HMC Port 1, although other network configurations are possible. You can attach a second HMC to HMC Port 2 of the server for redundancy (or vice versa). Figure 2-23 on page 60 shows a simple network configuration to enable the connection from HMC to server and to enable Dynamic LPAR operations. For more details about HMC and the possible network connections, refer to the *Hardware Management Console V7 Handbook*, SG24-7491.

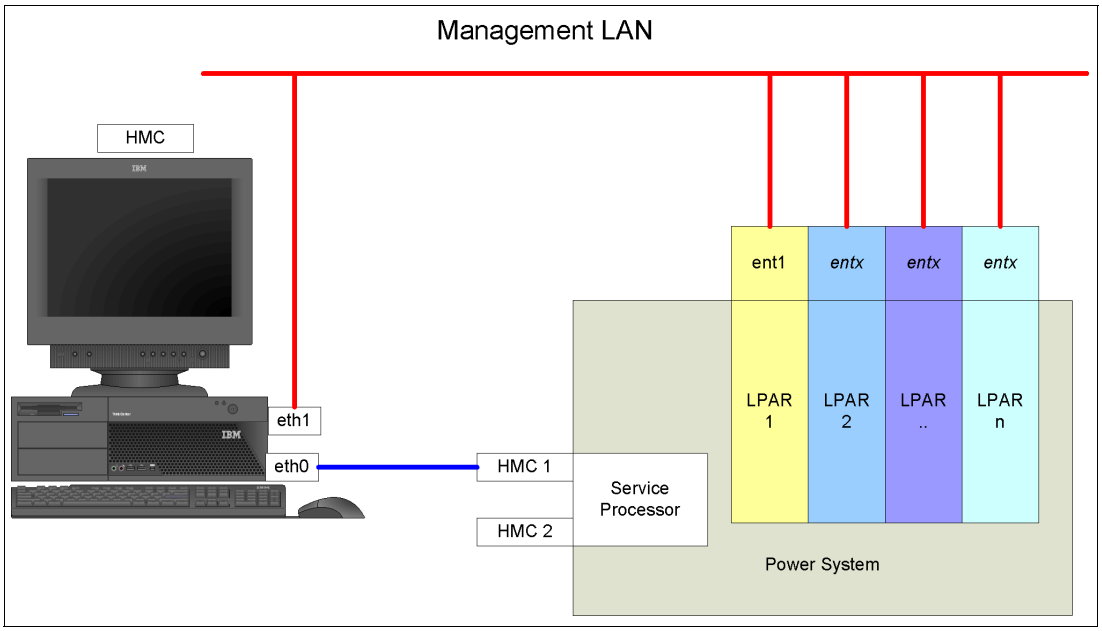


Figure 2-23 HMC to service processor and LPARs network connection

The default mechanism for allocation of the IP addresses for the service processor HMC ports is dynamic. The HMC can be configured as a DHCP server, providing the IP address at the time the managed server is powered on. If the service processor of the managed server does not receive DHCP reply before time-out, predefined IP addresses will setup on both ports. Static IP address allocation is also an option. You can configure the IP address of the service processor ports with a static IP address by using the Advanced System Management Interface (ASMI) menus.

Note: The service processor is used to monitor and manage the system hardware resources and devices. The service processor offers the following connections:

Two Ethernet 10/100 Mbps ports

- Both Ethernet ports are only visible to the service processor and can be used to attach the server to an HMC or to access the Advanced System Management Interface (ASMI) options from a client web browser, using the http server integrated into the service processor internal operating system.
- Both Ethernet ports have a default IP address
  - Service processor Eth0 or HMC1 port is configured as 169.254.2.147 with netmask 255.255.255.0
  - Service processor Eth1 or HMC2 port is configured as 169.254.3.147 with netmask 255.255.255.0

More information about the Service Processor can be found in 4.5.1, “Service processor” on page 116.

Functions performed by the HMC include:

- Creating and maintaining a multiple partition environment
- Displaying a virtual operating system session terminal for each partition
- Displaying a virtual operator panel of contents for each partition

- Detecting, reporting, and storing changes in hardware conditions
- Powering managed systems on and off
- Acting as a service focal point
- Generating or importing System Plans

The HMC provides both graphical and command line interface for all management tasks. Remote connection to the HMC using a web browser (as of HMC Version 7, previous versions required a special client program, called WebSM) or SSH are possible. The command line interface is also available by using the SSH secure shell connection to the HMC. It can be used by an external management system or a partition to perform HMC operations remotely.

2.12.1 High availability using the HMC

The HMC is an important hardware component. HACMP Version 5.4 high availability cluster software can be used to execute dynamic logical partitioning operations or activate additional resources (where available), thus becoming an integral part of the cluster.

If redundant HMC function is desired, the servers can be attached to two separate HMCs to address availability requirements. All HMCs must have the same level of Hardware Management Console Licensed Machine Code Version 7 (FC 0962) to manage POWER6 processor-based servers or an environment with a mixture of POWER5, POWER5+, and POWER6 processor-based servers. The HMCs provide a locking mechanism so that only one HMC at a time has write access to the service processor. Depending on your environment, you have multiple options to configure the network. Figure 2-24 shows one possible high available configuration.

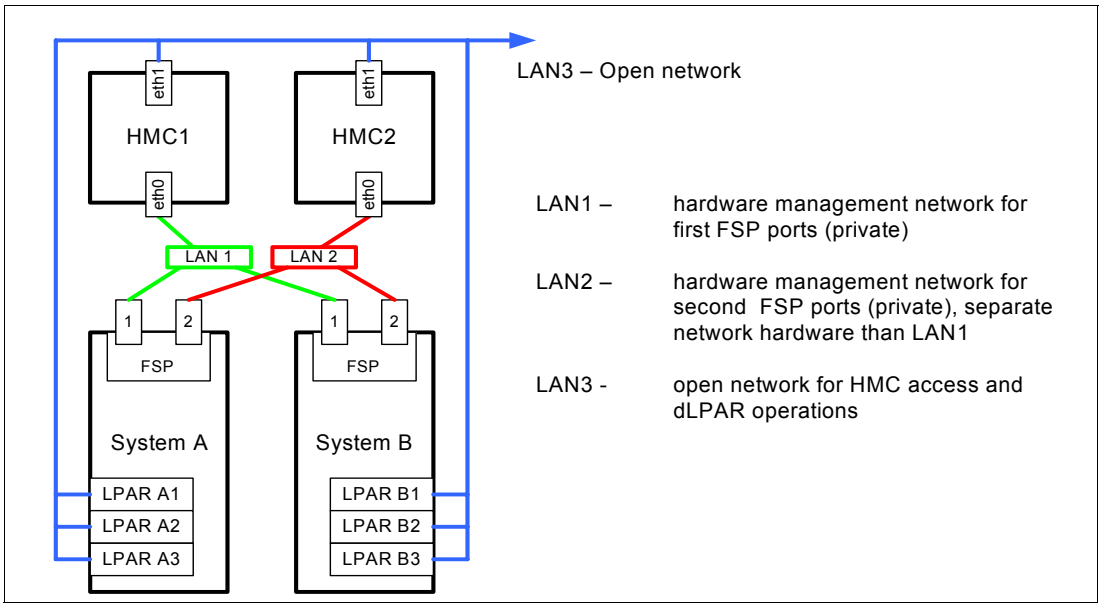


Figure 2-24 Highly available HMC and network architecture

Note that only hardware management networks (LAN1 and LAN2) are highly available on the above picture in order to keep simplicity. However, management network (LAN3) can be made highly available by using a similar concept and adding more Ethernet adapters to LPARs and HMCs.

Redundant Service Processor connectivity

Redundant Service Processor function for managing the service processors when one fails is supported on all systems that are operating with system firmware level FM320_xxx_xxx, or later. This support is available for configurations with two or more CEC enclosures. Redundant Service Processor function requires that the Hardware Management Console (HMC) be attached to the Service Interface Card in both CEC enclosure 1 and CEC enclosure 2. The Service Interface Card in these two enclosures must be connected using an external Power Control cable (FC 6006 or similar). Figure 2-25 shows a redundant HMC and redundant service processor connectivity configuration.

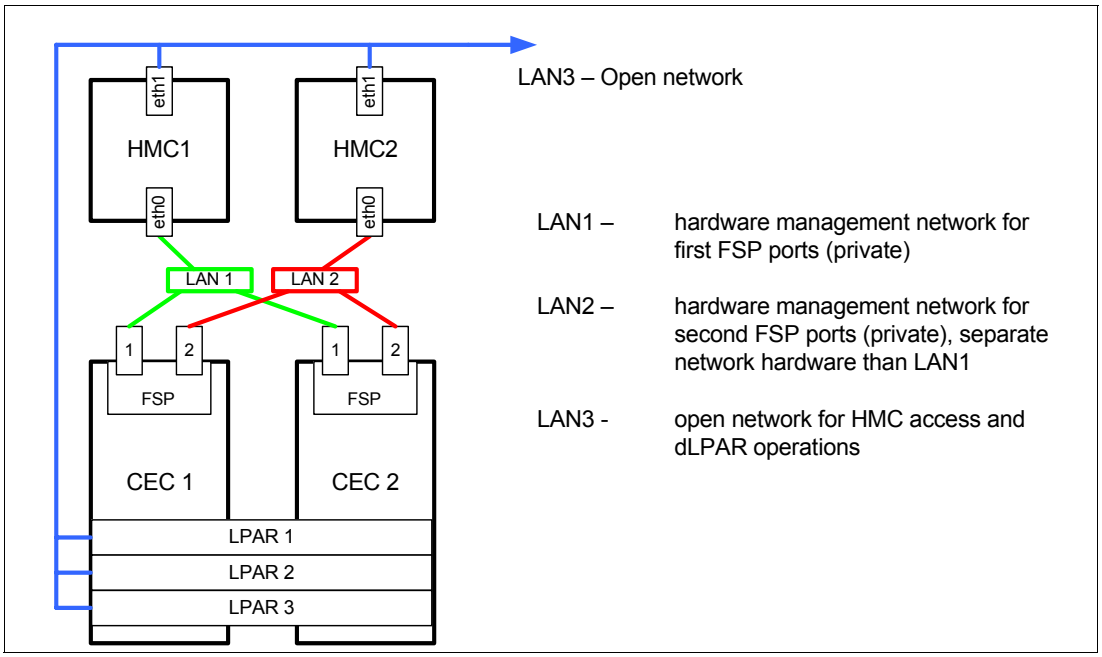


Figure 2-25 Redundant HMC connection and Redundant Service Processor configuration

In a configuration with multiple systems or HMC's, the customer is required to provide switches or hubs to connect each HMC to the appropriate Service Interface Cards in each system. One HMC should connect to the port labeled as HMC Port 1 on the first 2 CEC drawers of each system, and a second HMC should be attached to HMC Port 2 on the first 2 CEC drawers of each system. This provides redundancy both for the HMCs and the service processors.

For more details about redundant HMCs, refer to the *Hardware Management Console V7 Handbook*, SG24-7491.

2.12.2 Operating System Support

The POWER6-based IBM System 570 supports IBM AIX 5L Version 5.2, IBM AIX 5L Version 5.3, IBM AIX Version 6.1 and Linux distributions from SUSE and Red Hat.

Note: For specific technical support details, please refer to the support for IBM Web site:

`http://www.ibm.com/systems/p/os`

IBM AIX 5L

If installing AIX 5L on the 570, the following minimum requirements must be met:

- AIX 5L for POWER V5.3 with 5300-07 Technology Level (APAR IY99422), CD# LCD4-7463-09, DVD# LCD4-7544-05 or later
- AIX 5L for POWER V5.3 with 5300-06 Technology Level with Service Pack 4 (APAR IZ06992)
- AIX 5L for POWER V5.2 with 5200-10 Technology Level (APAR IY94898), CD# LCD4-1133-11

IBM periodically releases maintenance packages (service packs or technology levels) for the AIX5L operating system. These packages can be ordered on CD-ROM or downloaded from:

http://www-912.ibm.com/eserver/support/fixes/fixcentral/main/pseries/aix

The fixcentral Web site also provides information about how to obtain the CD-ROM.

You can also get individual operating system fixes and information about obtaining AIX 5L service at this site. From AIX 5L V5.3 the Service Update Management Assistant, which helps the administrator to automate the task of checking and downloading operating system downloads, is part of the base operating system. For more information about the **suma** command functionality, refer to:

http://www14.software.ibm.com/webapp/set2/sas/f/suma/home.html

AIX 5L is supported on the System p servers in partitions with dedicated processors (LPARs), and shared-processor partitions (micro-partitions). When combined with one of the PowerVM features, AIX 5L Version 5.3 can make use of all the existing and new virtualization features such as micro-partitions, virtual I/O, virtual LAN, and PowerVM Live Partition Mobility, to name a few.

IBM AIX V6.1

IBM is making available a new version of AIX, AIX V6.1 which will include significant new capabilities for virtualization, security features, continuous availability features and manageability. AIX V6.1 is the first generally available version of AIX V6.

AIX V6.1 features include support for:

- PowerVM AIX 6 Workload Partitions (WPAR) - software based virtualization
- Live Application Mobility - with the IBM PowerVM AIX 6 Workload Partitions Manager for AIX (5765-WPM)
- 64-bit Kernel for higher scalability and performance
- Dynamic logical partitioning and Micro-Partitioning support
- Support for Multiple Shared-Processor Pools
- Trusted AIX - MultiLevel, compartmentalized security
- Integrated Role Based Access Control
- Encrypting JFS2 file system
- Kernel exploitation of POWER6 Storage Keys for greater reliability
- Robust journaled file system and Logical Volume Manager (LVM) software including integrated file system snapshot
- Tools for managing the systems environment -- System Management

- Interface Tool (SMIT) and the IBM Systems Director Console for AIX

Linux for System p systems

Linux is an open source operating system that runs on numerous platforms from embedded systems to mainframe computers. It provides a UNIX-like implementation across many computer architectures.

This section discusses two brands of Linux to be run in partitions. The supported versions of Linux on System p servers are:

- Novell SUSE Linux Enterprise Server V10 SP1 for POWER Systems or later
- Red Hat Enterprise Linux Advanced Server V4.5 for POWER or later
- Red Hat Enterprise Linux Advanced Server V5.1 for Power or later

The PowerVM features are supported in Version 2.6.9 and above of the Linux kernel. The commercially available latest distributions from Red Hat, Inc. (RHEL AS 5) and Novell SUSE Linux (SLES 10) support the IBM system p 64-bit architectures and are based on this 2.6 kernel series.

Clients wishing to configure Linux partitions in virtualized System p systems should consider the following:

- Not all devices and features supported by the AIX operating system are supported in logical partitions running the Linux operating system.
- Linux operating system licenses are ordered separately from the hardware. Clients can acquire Linux operating system licenses from IBM, to be included with their System 570 or from other Linux distributors.

For information about the features and external devices supported by Linux refer to:

http://www-03.ibm.com/systems/p/os/linux/index.html

For information about SUSE Linux Enterprise Server 10, refer to:

http://www.novell.com/products/server

For information about Red Hat Enterprise Linux Advanced Server 5, refer to:

http://www.redhat.com/rhel/features

Supported virtualization features

SLES 10, RHEL AS 4.5 and RHEL AS 5 support the following virtualization features:

- Virtual SCSI, including for the boot device
- Shared-processor partitions and virtual processors, capped and uncapped
- Dedicated-processor partitions
- Dynamic reconfiguration of processors
- Virtual Ethernet, including connections through the Shared Ethernet Adapter in the Virtual I/O Server to a physical Ethernet connection
- Simultaneous multithreading (SMT)

SLES 10, RHEL AS 4.5, and RHEL AS 5 do not support the following:

- Dynamic reconfiguration of memory
- Dynamic reconfiguration of I/O slot

Note: IBM only supports the Linux systems of clients with a SupportLine contract covering Linux. Otherwise, contact the Linux distributor for support.

i5/OS

At the time of writing i5/OS® is not supported.

2.13 Service information

The 570 is not a client setup server (CSU). Therefore, the IBM service representative completes the system installation.

2.13.1 Touch point colors

Blue (IBM blue) or terra-cotta (orange) on a component indicates a touch point (for electronic parts) where you can grip the hardware to remove it from or to install it into the system, to open or to close a latch, and so on. IBM defines the touch point colors as follows:

Blue	This requires a shutdown of the system before the task can be performed, for example, installing additional processors contained in the second processor book.
Terra-cotta	The system can remain powered on while this task is being performed. Keep in mind that some tasks might require that other steps have to be performed first. One example is deconfiguring a physical volume in the operating system before removing a disk from a 4-pack disk enclosure of the server.
Blue and terra-cotta	Terra-cotta takes precedence over this color combination, and the rules for a terra-cotta-only touch point apply.

Important: It is important to adhere to the touch point colors on the system. Not doing so can compromise your safety and damage the system.

2.13.2 Operator Panel

The service processor provides an interface to the control panel that is used to display server status and diagnostic information. See Figure 2-26 on page 66 for operator control panel physical details and buttons.

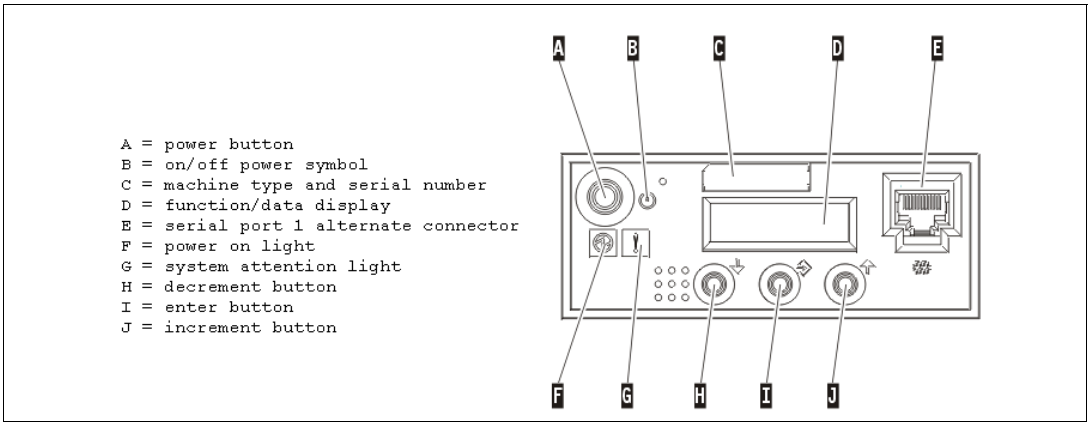


Figure 2-26 Operator control panel physical details and button

Note: For servers managed by the HMC, use it to perform control panel functions.

Primary control panel functions

The primary control panel functions are defined as functions 01 to 20, including options to view and manipulate IPL modes, server operating modes, IPL speed, and IPL type.

The following list describes the primary functions:

- Function 01: Display the selected IPL type, system operating mode, and IPL speed
- Function 02: Select the IPL type, IPL speed override, and system operating mode
- Function 03: Start IPL
- Function 04: Lamp Test
- Function 05: Reserved
- Function 06: Reserved
- Function 07: SPCN functions
- Function 08: Fast Power Off
- Functions 09 to 10: Reserved
- Functions 11 to 19: System Reference Code
- Function 20: System type, model, feature code, and IPL type

All the functions mentioned are accessible using the Advanced System Management Interface (ASMI), HMC, or the control panel.

Extended control panel functions

The extended control panel functions consist of two major groups:

- Functions 21 through 49, which are available when you select Manual mode from Function 02.
- Support service representative Functions 50 through 99, which are available when you select Manual mode from Function 02, then select and enter the client service switch 1 (Function 25), followed by service switch 2 (Function 26).

Function 30 – CEC SP IP address and location

Function 30 is one of the Extended control panel functions and is only available when Manual mode is selected. This function can be used to display the central electronic complex (CEC) Service Processor IP address and location segment. The Table 2-21 shows an example of how to use the Function 03.

Table 2-21 CEC SP IP address and location

Information on operator panel	Action or description
3 0	Use the increment or decrement buttons to scroll to Function 30.
3 0 * *	Press Enter to enter sub-function mode.
3 0 0 0	Use the increment or decrement buttons to select an IP address: 0 0 = Service Processor ETH0 or HMC1 port 0 1 = Service Processor ETH1 or HMC2 port
S P A : E T H 0 : _ _ _ T 5 1 9 2 . 1 6 8 . 2 . 1 4 7	Press Enter to display the selected IP address.
3 0 * *	Use the increment or decrement buttons to select sub-function exit.
3 0	Press Enter to exit sub-function mode.

2.14 System firmware

Server firmware is the part of the Licensed Internal Code that enables hardware, such as the service processor. Depending on your service environment, you can download, install, and manage your server firmware fixes using different interfaces and methods, including the HMC, or by using functions specific to your operating system.

Note: Normally, installing the server firmware fixes through the operating system is a nonconcurrent process.

Temporary and permanent firmware sides

The service processor maintains two copies of the server firmware:

- One copy is considered the permanent or backup copy and is stored on the permanent side, sometimes referred to as the *p* side.
- The other copy is considered the installed or temporary copy and is stored on the temporary side, sometimes referred to as the *t* side. We recommend that you start and run the server from the temporary side.

The copy actually booted from is called the activated level, sometimes referred to as *b*.

Note: The default value, from which the system boots, is temporary.

The following examples are the output of the **lsmcode** command for AIX and Linux, showing the firmware levels as they are displayed in the outputs.

- AIX:

```
The current permanent system firmware image is SF220_005.
The current temporary system firmware image is SF220_006.
The system is currently booted from the temporary image.
```

- Linux:

```
system:SF220_006 (t) SF220_005 (p) SF220_006 (b)
```

When you install a server firmware fix, it is installed on the temporary side.

Note: The following points are of special interest:

- The server firmware fix is installed on the temporary side only after the existing contents of the temporary side are permanently installed on the permanent side (the service processor performs this process automatically when you install a server firmware fix).
- If you want to preserve the contents of the permanent side, you need to remove the current level of firmware (copy the contents of the permanent side to the temporary side) before you install the fix.
- However, if you get your fixes using the Advanced features on the HMC interface and you indicate that you do not want the service processor to automatically accept the firmware level, the contents of the temporary side are not automatically installed on the permanent side. In this situation, you do not need to remove the current level of firmware to preserve the contents of the permanent side before you install the fix.

You might want to use the new level of firmware for a period of time to verify that it works correctly. When you are sure that the new level of firmware works correctly, you can permanently install the server firmware fix. When you permanently install a server firmware fix, you copy the temporary firmware level from the temporary side to the permanent side.

Conversely, if you decide that you do not want to keep the new level of server firmware, you can remove the current level of firmware. When you remove the current level of firmware, you copy the firmware level that is currently installed on the permanent side from the permanent side to the temporary side.

System firmware download site

For the system firmware download site for the 570, go to:

http://www14.software.ibm.com/webapp/set2/firmware/gjsn

In the main area of the firmware download site, select the correct machine type and model. The 570 machine type and model is 9117-MMA (see Figure 2-27 on page 69)

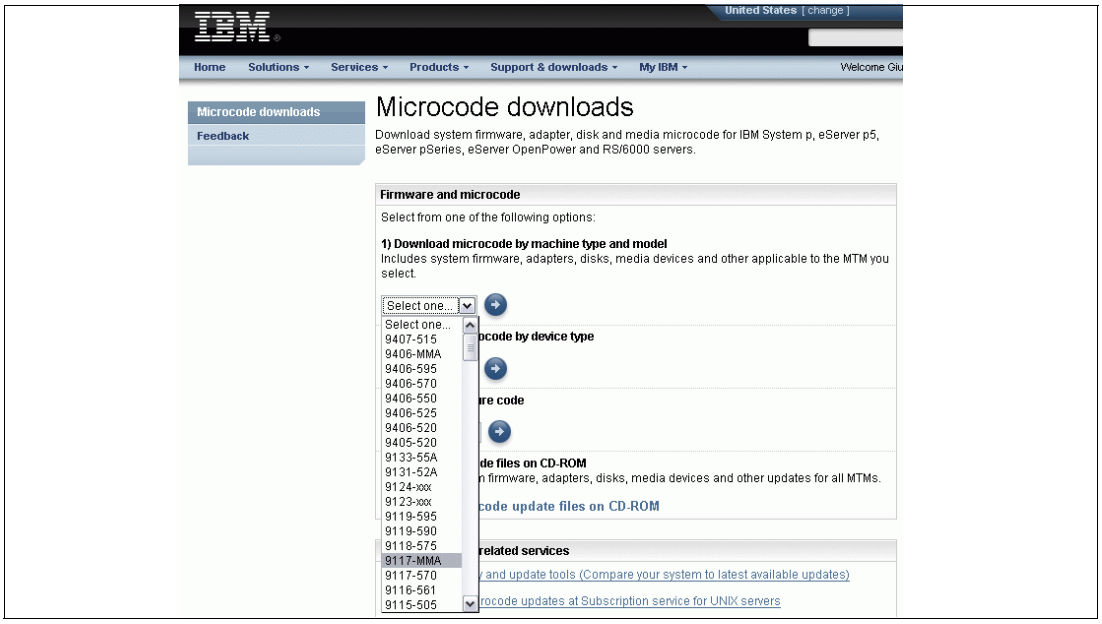


Figure 2-27 IBM Microcode downloads site

Receive server firmware fixes using an HMC

If you use an HMC to manage your server and you periodically configure several partitions on the server, you need to download and install fixes for your server and power subsystem firmware.

How you get the fix depends on whether the HMC or server is connected to the Internet:

- The HMC or server is connected to the Internet.

  There are several repository locations from which you can download the fixes using the HMC. For example, you can download the fixes from your service provider's Web site or support system, from optical media that you order from your service provider, or from an FTP server on which you previously placed the fixes.

- Neither the HMC nor your server is connected to the Internet (server firmware only).

  You need to download your new server firmware level to a CD-ROM media or FTP server.

For both of these options, you can use the interface on the HMC to install the firmware fix (from one of the repository locations or from the optical media). The Change Internal Code wizard on the HMC provides a step-by-step process for you to perform the procedure to install the fix. Perform these steps:

1. Ensure that you have a connection to the service provider (if you have an Internet connection from the HMC or server).
2. Determine the available levels of server and power subsystem firmware.
3. Create the optical media (if you do not have an Internet connection from the HMC or server).
4. Use the Change Internal Code wizard to update your server and power subsystem firmware.
5. Verify that the fix installed successfully.

For a detailed description of each task, select **System p information, support, and troubleshooting** → **Fixes and upgrades** → **Getting fixes and upgrades** from the IBM Systems Hardware Information Center Web site at:

http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp?lang=en

Receive server firmware fixes without an HMC

Periodically, you need to install fixes for your server firmware. If you do not use an HMC to manage your server, you must get your fixes through your operating system. In this situation, you can get server firmware fixes through the operating system regardless of whether your operating system is AIX or Linux.

To do this, complete the following tasks:

1. Determine the existing level of server firmware using the **lsmcode** command.
2. Determine the available levels of server firmware.
3. Get the server firmware.
4. Install the server firmware fix to the temporary side.
5. Verify that the server firmware fix installed successfully.
6. Install the server firmware fix permanently (optional).

Note: To view existing levels of server firmware using the **lsmcode** command, you need to have the following service tools installed on your server:

- AIX

  You must have AIX diagnostics installed on your server to perform this task. AIX diagnostics are installed when you install AIX on your server. However, it is possible to deselect the diagnostics. Therefore, you need to ensure that the online AIX diagnostics are installed before proceeding with this task.

- Linux
  - Platform Enablement Library: librtas-*nnnnn*.rpm
  - Service Aids: ppc64-utils-*nnnnn*.rpm
  - Hardware Inventory: lsvpd-*nnnnn*.rpm

  Where *nnnnn* represents a specific version of the RPM file.

  If you do not have the service tools on your server, you can download them at the following Web site:

  http://www14.software.ibm.com/webapp/set2/sas/flopdiags/home.html

2.14.1 Service processor

The service processor is an embedded controller running the service processor internal operating system. The service processor operating system contains specific programs and device drivers for the service processor hardware. The host interface is a 32-bit PCI-X interface connected to the Enhanced I/O Controller.

Service processor is used to monitor and manage the system hardware resources and devices. The service processor offers the following connections:

Two Ethernet 10/100 Mbps ports

- Both Ethernet ports are only visible to the service processor and can be used to attach the p5-570 to a HMC or to access the Advanced System Management Interface (ASMI)

options from a client Web browser, using the HTTP server integrated into the service processor internal operating system.

- Both Ethernet ports have a default IP address:
  - Service processor Eth0 or HMC1 port is configured as 169.254.2.147.
  - Service processor Eth1 or HMC2 port is configured as 169.254.3.147.

2.14.2 Redundant service processor

A Service Interface card is required to be installed in every drawer. The card in the top drawer provides the Service Processor function. Redundant Service Processor function for managing the service processors when one fails is supported on all systems that are operating with system firmware level FM320_xxx_xxx, or later. This support is available for configurations with two or more CEC enclosures. The card in the second drawer provides a Service Processor on standby capable of taking over the Service Processor function from the primary drawer.

The SP Flash in the second drawer will be updated whenever an update is made to the SP Flash in the top drawer. The Service Interface cards in drawers 3 and 4 do not use the Service Processor function, and the FLASH code is not updated. Therefore Service Interface cards from drawers 3 or 4 must NOT be moved into drawers 1 or 2.

Redundant Service Processor function requires that the Hardware Management Console (HMC) be attached to the Service Interface Card in both CEC enclosure 1 and CEC enclosure 2. The Service Interface Card in these two enclosures must be connected using an external Power Control cable (FC 6006 or similar).

http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp?topic=/iphae/planredundantfsp.htm

2.14.3 Hardware management user interfaces

This section provides a brief overview of the different 570 hardware management user interfaces available.

Advanced System Management Interface

The Advanced System Management Interface (ASMI) is the interface to the service processor that enables you to set flags that affect the operation of the server, such as auto power restart, and to view information about the server, such as the error log and vital product data.

This interface is accessible using a Web browser on a client system that is connected directly to the service processor (in this case, a standard Ethernet cable or a crossed cable can be both used) or through an Ethernet network. Using the *network configuration menu*, the ASMI enables the ability to change the service processor IP addresses or to apply some security policies and avoid the access from undesired IP addresses or range. The ASMI can also be accessed using a terminal attached to the system service processor ports on the server, if the server is not HMC managed. The service processor and the ASMI are standard on all IBM System p servers.

You might be able to use the service processor's default settings. In that case, accessing the ASMI is not necessary.

Accessing the ASMI using a Web browser

The Web interface to the Advanced System Management Interface is accessible through, at the time of writing, Microsoft® Internet Explorer® 6.0, Netscape 7.1, Mozilla Firefox, or

Opera 7.23 running on a PC or mobile computer connected to the service processor. The Web interface is available during all phases of system operation, including the initial program load and runtime. However, some of the menu options in the Web interface are unavailable during IPL or runtime to prevent usage or ownership conflicts if the system resources are in use during that phase.

Accessing the ASMI using an ASCII console

The Advanced System Management Interface on an ASCII console supports a subset of the functions provided by the Web interface and is available only when the system is in the platform standby state. The ASMI on an ASCII console is not available during some phases of system operation, such as the initial program load and runtime.

Accessing the ASMI using an HMC

To access the Advanced System Management Interface using the Hardware Management Console, complete the following steps:

1. Open Systems Management from the navigation pane.
2. From the work pane, select one or more managed systems to work with.
3. From the System Management tasks list, select Operations.
4. From the Operations task list, select Advanced System Management (ASM).

System Management Services

Use the System Management Services (SMS) menus to view information about your system or partition and to perform tasks, such as changing the boot list or setting the network parameters.

To start System Management Services, perform the following steps:

1. For a server that is connected to an HMC, use the HMC to restart the server or partition.

   If the server is not connected to an HMC, stop the system, and then restart the server by pressing the power button on the control panel.
2. For a partitioned server, watch the virtual terminal window on the HMC.

   For a full server partition, watch the firmware console.
3. Look for the POST[5] indicators (memory, keyboard, network, SCSI, and speaker) that appear across the bottom of the screen. Press the numeric 1 key after the word `keyboard` appears and before the word `speaker` appears.

The SMS menus is useful to defining the operating system installation method, choosing the installation boot device, or setting the boot device priority list for a full managed server or a logical partition. In the case of a network boot, SMS menus are provided to set up the network parameters and network adapter IP address.

HMC

The Hardware Management Console is a system that controls managed systems, including IBM System p5 and p6 hardware, and logical partitions. To provide flexibility and availability, there are different ways to implement HMCs.

Web-based System Manager Remote Client

The Web-based System Manager Remote Client is an application that is usually installed on a PC and can be downloaded directly from an installed HMC. When an HMC is installed and

[5] POST stands for power-on-self-test.

HMC Ethernet IP addresses have been assigned, it is possible to download the Web-based System Manager Remote Client from a web browser, using the following URL:

```
http://HMC_IP_address/remote_client.html
```

You can then use the PC to access other HMCs remotely. Web-based System Manager Remote Clients can be present in private and open networks. You can perform most management tasks using the Web-based System Manager Remote Client. The remote HMC and the Web-based System Manager Remote Client allow you the flexibility to access your managed systems (including HMCs) from multiple locations using multiple HMCs.

For more detailed information about the use of the HMC, refer to the IBM Systems Hardware Information Center.

Open Firmware

A System p6 server has one instance of Open Firmware both when in the partitioned environment and when running as a full system partition. Open Firmware has access to all devices and data in the server. Open Firmware is started when the server goes through a power-on reset. Open Firmware, which runs in addition to the POWER Hypervisor in a partitioned environment, runs in two modes: global and partition. Each mode of Open Firmware shares the same firmware binary that is stored in the flash memory.

In a partitioned environment, Open Firmware runs on top of the global Open Firmware instance. The partition Open Firmware is started when a partition is activated. Each partition has its own instance of Open Firmware and has access to all the devices assigned to that partition. However, each instance of Open Firmware has no access to devices outside of the partition in which it runs. Partition firmware resides within the partition memory and is replaced when AIX or Linux takes control. Partition firmware is needed only for the time that is necessary to load AIX or Linux into the partition server memory.

The global Open Firmware environment includes the partition manager component. That component is an application in the global Open Firmware that establishes partitions and their corresponding resources (such as CPU, memory, and I/O slots), which are defined in partition profiles. The partition manager manages the operational partitioning transactions. It responds to commands from the service processor external command interface that originates in the application running on the HMC. The ASMI can be accessed during boot time or using the ASMI and selecting the boot to Open Firmware prompt.

For more information about Open Firmware, refer to *Partitioning Implementations for IBM* eServer *Partitioning Implementations for IBM @server p5 Servers*, SG24-7039, which is available at:

http://www.redbooks.ibm.com/abstracts/sg247039.html

3

Virtualization

As you look for ways to maximize the return on your IT infrastructure investments, consolidating workloads becomes an attractive proposition.

IBM Power Systems combined with PowerVM technology are designed to help you consolidate and simplify your IT environment. Key capabilities include:

- Improve server utilization and sharing I/O resources to reduce total cost of ownership and make better use of IT assets.
- Improve business responsiveness and operational speed by dynamically re-allocating resources to applications as needed — to better match changing business needs or handle unexpected changes in demand.
- Simplify IT infrastructure management by making workloads independent of hardware resources, thereby enabling you to make business-driven policies to deliver resources based on time, cost and service-level requirements.

This chapter discusses the virtualization technologies and features on IBM Power Systems:

- POWER Hypervisor
- Logical Partitions
- Dynamic Logical Partitioning
- Shared Processor Pool
- PowerVM

© Copyright IBM Corp. 2008. All rights reserved.

3.1 POWER Hypervisor

Combined with features designed into the POWER6 processors, the POWER Hypervisor delivers functions that enable other system technologies, including logical partitioning technology, virtualized processors, IEEE VLAN compatible virtual switch, virtual SCSI adapters, and virtual consoles. The POWER Hypervisor is a basic component of the system's firmware and offers the following functions:

- Provides an abstraction between the physical hardware resources and the logical partitions that use them.
- Enforces partition integrity by providing a security layer between logical partitions.
- Controls the dispatch of virtual processors to physical processors (see 3.2.3, "Processing mode" on page 79).
- Saves and restores all processor state information during a logical processor context switch.
- Controls hardware I/O interrupt management facilities for logical partitions.
- Provides virtual LAN channels between logical partitions that help to reduce the need for physical Ethernet adapters for inter-partition communication.
- Monitors the Service Processor and will perform a reset/reload if it detects the loss of the Service Processor, notifying the operating system if the problem is not corrected.

The POWER Hypervisor is always active, regardless of the system configuration and also when not connected to the HMC. It requires memory to support the resource assignment to the logical partitions on the server. The amount of memory required by the POWER Hypervisor firmware varies according to several factors. Factors influencing the POWER Hypervisor memory requirements include the following:

- Number of logical partitions.
- Number of physical and virtual I/O devices used by the logical partitions.
- Maximum memory values given to the logical partitions.

The minimum amount of physical memory to create a partition is the size of the system's Logical Memory Block (LMB). The default LMB size varies according to the amount of memory configured in the CEC as shown in Table 3-1.

Table 3-1 Configurable CEC memory-to-default Logical Memory Block size

Configurable CEC memory	Default Logical Memory Block
Less than 4 GB	16 MB
Greater than 4 GB up to 8 GB	32 MB
Greater than 8 GB up to 16 GB	64 MB
Greater than 16 GB up to 32 GB	128 MB
Greater than 32 GB	256 MB

But in most cases, the actual requirements and recommendations are between 256 MB and 512 MB for AIX, Red Hat, and Novell SUSE Linux. Physical memory is assigned to partitions in increments of Logical Memory Block (LMB).

The POWER Hypervisor provides the following types of virtual I/O adapters:

- Virtual SCSI

- Virtual Ethernet
- Virtual (TTY) console

Virtual SCSI

The POWER Hypervisor provides a virtual SCSI mechanism for virtualization of storage devices (a special logical partition to install the Virtual I/O Server is required to use this feature, as described in 3.3.2, “Virtual I/O Server” on page 81). The storage virtualization is accomplished using two, paired, adapters: a virtual SCSI server adapter and a virtual SCSI client adapter. Only the Virtual I/O Server partition can define virtual SCSI server adapters, other partitions are *client* partitions. The Virtual I/O Server is available with the optional PowerVM Edition features.

Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions on the same server to use a fast and secure communication without any need for physical interconnection. The virtual Ethernet allows a transmission speed in the range of 1 to 3 Gbps. depending on the MTU[1] size and CPU entitlement. Virtual Ethernet support starts with AIX 5L Version 5.3, or appropriate level of Linux supporting Virtual Ethernet devices (see chapter 3.3.7, “Operating System support for PowerVM” on page 90). The virtual Ethernet is part of the base system configuration.

Virtual Ethernet has the following major features:

- The virtual Ethernet adapters can be used for both IPv4 and IPv6 communication and can transmit packets with a size up to 65408 bytes. Therefore, the maximum MTU for the corresponding interface can be up to 65394 (65390 if VLAN tagging is used).
- The POWER Hypervisor presents itself to partitions as a virtual 802.1Q compliant switch. The maximum number of VLANs is 4096. Virtual Ethernet adapters can be configured as either untagged or tagged (following the IEEE 802.1Q VLAN standard).
- A partition supports 256 virtual Ethernet adapters. Besides a default port VLAN ID, the number of additional VLAN ID values that can be assigned per Virtual Ethernet adapter is 20, which implies that each Virtual Ethernet adapter can be used to access 21 virtual networks.
- Each partition operating system detects the virtual local area network (VLAN) switch as an Ethernet adapter without the physical link properties and asynchronous data transmit operations.

Any virtual Ethernet can also have connectivity outside of the server if a layer-2 bridge to a physical Ethernet adapter is set in one Virtual I/O server partition (see 3.3.2, “Virtual I/O Server” on page 81 for more details about shared Ethernet). Also known as Shared Ethernet Adapter.

Note: Virtual Ethernet is based on the IEEE 802.1Q VLAN standard. No physical I/O adapter is required when creating a VLAN connection between partitions, and no access to an outside network is required.

Virtual (TTY) console

Each partition needs to have access to a system console. Tasks such as operating system installation, network setup, and some problem analysis activities require a dedicated system console. The POWER Hypervisor provides the virtual console using a virtual TTY or serial

[1] Maximum transmission unit

adapter and a set of Hypervisor calls to operate on them. Virtual TTY does not require the purchase of any additional features or software such as the PowerVM Edition features.

Depending on the system configuration, the operating system console can be provided by the Hardware Management Console virtual TTY, IVM virtual TTY, or from a terminal emulator that is connected to a system port.

3.2 Logical partitioning

Logical partitions (LPARs) and virtualization increase utilization of system resources and add a new level of configuration possibilities. This section provides details and configuration specifications about this topic.

3.2.1 Dynamic logical partitioning

Logical partitioning (LPAR) was introduced with the POWER4™ processor-based product line and the AIX 5L Version 5.1 operating system. This technology offered the capability to divide a pSeries system into separate logical systems, allowing each LPAR to run an operating environment on dedicated attached devices, such as processors, memory, and I/O components.

Later, dynamic logical partitioning increased the flexibility, allowing selected system resources, such as processors, memory, and I/O components, to be added and deleted from logical partitions while they are executing. AIX 5L Version 5.2, with all the necessary enhancements to enable dynamic LPAR, was introduced in 2002. The ability to reconfigure dynamic LPARs encourages system administrators to dynamically redefine all available system resources to reach the optimum capacity for each defined dynamic LPAR.

3.2.2 Micro-Partitioning

Micro-Partitioning technology allows you to allocate fractions of processors to a logical partition. This technology was introduced with POWER5 processor-based systems. A logical partition using fractions of processors is also known as a Shared Processor Partition or Micro-Partition. Micro-Partitions run over a set of processors called Shared Processor Pool. And virtual processors are used to let the operating system manage the fractions of processing power assigned to the logical partition. From an operating system perspective, a virtual processor cannot be distinguished from a physical processor, unless the operating system has been enhanced to be made aware of the difference. Physical processors are abstracted into virtual processors that are available to partitions. The meaning of the term *physical processor* in this section is a *processor core*. For example, in a 2-core server there are two physical processors.

When defining a shared processor partition, several options have to be defined:

- The minimum, desired, and maximum processing units. Processing units are defined as processing power, or the fraction of time the partition is dispatched on physical processors. Processing units define the capacity entitlement of the partition.
- The shared processor pool. Pick one from the list with the names of each configured shared processor pool. This list also displays the pool ID of each configured shared processor pool in parentheses. If the name of the desired shared processor pool is not available here, you must first configure the desired shared processor pool using the Shared Processor Pool Management window. Shared processor partitions use the default shared processor pool called DefaultPool by default.

- Select whether the partition will be able or not to access extra processing power to “fill up” its virtual processors above its capacity entitlement. Selecting either to capp or uncapp your partition. If there is spare processing power available in the shared processor pool or other partitions are not using their entitlement, an uncapped partition can use additional processing units if its entitlement is not enough to satisfy its application processing demand.
- The weight (preference) in the case of an uncapped partition.
- The minimum, desired, and maximum number of virtual processors.

The POWER Hypervisor calculates partition’s processing power based on minimum, desired, and maximum values, processing mode and also based on other active partitions’ requirements. The actual entitlement is never smaller than the processing units desired value but can exceed that value in the case of an uncapped partition and up to the number of virtual processors allocated.

A partition can be defined with a processor capacity as small as 0.10 processing units. This represents 1/10th of a physical processor. Each physical processor can be shared by up to 10 shared processor partitions and the partition’s entitlement can be incremented fractionally by as little as 1/100th of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC or Integrated Virtualization Management.

This system supports up to a 16-core configuration, therefore up to sixteen dedicated partitions, or up to 160 micro-partitions, can be created. It is important to point out that the maximums stated are supported by the hardware, but the practical limits depend on the application workload demands.

Additional information on virtual processors:

- A virtual processor can be either running (dispatched) on a physical processor or standby waiting for a physical processor to became available.
- Virtual processors do not introduce any additional abstraction level; they really are only a dispatch entity. When running on a physical processor, virtual processors run at the same speed as the physical processor.
- Each partition’s profile defines CPU entitlement that determines how much processing power any given partition should receive. The total sum of CPU entitlement of all partitions cannot exceed the number of available physical processors in a shared processor pool.
- The number of virtual processors can be changed dynamically through a dynamic LPAR operation.

3.2.3 Processing mode

When you create a logical partition you can assign entire processors for dedicated use, or you can assign partial processor units from a shared processor pool. This setting will define the processing mode of the logical partition. Figure 3-1 on page 80 shows a diagram of the concepts discussed in this section.

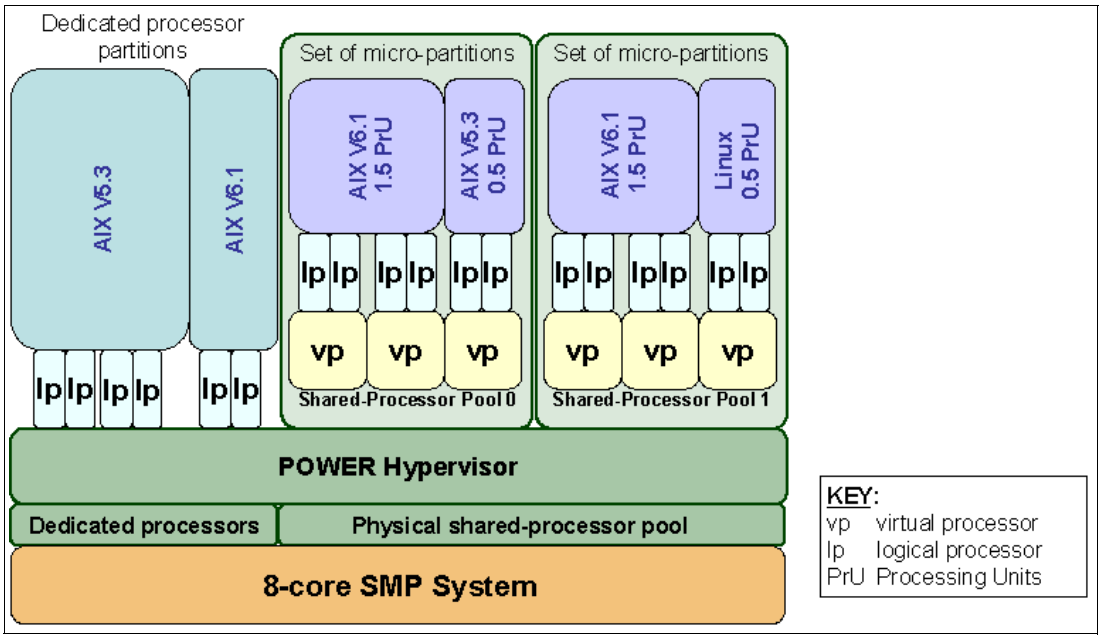


Figure 3-1 Logical partitioning concepts

Dedicated mode

In dedicated mode, physical processors are assigned as a whole to partitions. The simultaneous multithreading feature in the POWER6 processor core allows the core to execute instructions from two independent software threads simultaneously. To support this feature we use the concept of *logical processors*. The operating system (AIX or Linux) sees one physical processor as two logical processors if the simultaneous multithreading feature is on. It can be turned off and on dynamically while the operating system is executing (for AIX, use the **smtctl** command). If simultaneous multithreading is off, then each physical processor is presented as one logical processor and thus only one thread at a time is executed on the physical processor.

Shared dedicated mode

On POWER6 servers, you can configure dedicated partitions to become processor donors for idle processors they own. Allowing for the donation of spare CPU cycles from dedicated processor partitions to a Shared Processor Pool. The dedicated partition maintains absolute priority for dedicated CPU cycles. Enabling this feature may help to increase system utilization, without compromising the computing power for critical workloads in a dedicated processor.

Shared mode

In shared mode, logical partitions use virtual processors to access fractions of physical processors. Shared partitions can define any number of virtual processors (maximum number is 10 times the number of processing units assigned to the partition). From the POWER Hypervisor point of view, virtual processors represent dispatching objects. The POWER Hypervisor dispatches virtual processors to physical processors according to partition's processing units entitlement. One Processing Unit represents one physical processor's processing capacity. At the end of the POWER Hypervisor's dispatch cycle (10 ms), all partitions should receive total CPU time equal to their processing units entitlement. The logical processors are defined on top of virtual processors. So, even with a virtual processor, the concept of logical processor exists and the number of logical processor depends whether the simultaneous multithreading is turned on or off.

3.3 PowerVM

The PowerVM platform is the family of technologies, capabilities and offerings that deliver industry-leading virtualization on this 570. It is the new umbrella branding term for Power Systems Virtualization (Logical Partitioning, Micro-Partitioning, Hypervisor, Virtual I/O Server, Advanced Power Virtualization, Live Partition Mobility, Workload Partitions, etc.). As with Advanced Power Virtualization in the past, PowerVM is a combination of hardware enablement and value-added software. Section 3.3.1, “PowerVM editions” on page 81 discusses the licensed features of each of the 2 different editions of PowerVM.

3.3.1 PowerVM editions

This section provides information about the virtualization capabilities of the PowerVM Standard Edition and Enterprise Edition which are available on this system. Upgrading from the PowerVM Standard Edition to the PowerVM Express Edition is possible and is completely undisruptive. The upgrade doesn’t even require the installation of additional software, the customer just has to enter a key code in the hypervisor in order to unlock the next level of function.

Table 3-2, outlines the functional elements of both of the PowerVM editions.

Table 3-2 PowerVM capabilities

PowerVM capability	PowerVM Standard Edition (FC 7942)	PowerVM Enterprise Edition (FC 7995)
Micro-partitions	Yes	Yes
Virtual I/O Server	Yes	Yes
Shared Dedicated Capacity	Yes	Yes
Multiple Shared-Processor Pools	Yes	Yes
Lx86	Yes	Yes
Live Partition Mobility	No	Yes
Maximum # Logical Partitions	Up to 10 per core	Up to 10 per core

For more information about the different PowerVM editions please refer to *PowerVM Virtualization on IBM System p Introduction and Configuration*, SG24-7940.

Note The 570 has to be managed with the Hardware Management Console.

3.3.2 Virtual I/O Server

The Virtual I/O Server is part of all PowerVM Editions. It is a special purpose partition that allows the sharing of physical resources between logical partitions to allow more efficient utilization (for example consolidation). In this case the Virtual I/O Server owns the physical resources (SCSI, Fibre Channel, network adapters, and optical devices) and allows client partitions to share access to them, thus minimizing the number of physical adapters in the system. The Virtual I/O Server eliminates the requirement that every partition owns a dedicated network adapter, disk adapter, and disk drive. The Virtual I/O Server supports OpenSSH for secure remote logins. It also provides a firewall for limiting access by ports,

network services and IP addresses. Figure 3-2 shows an overview of a Virtual I/O Server configuration.

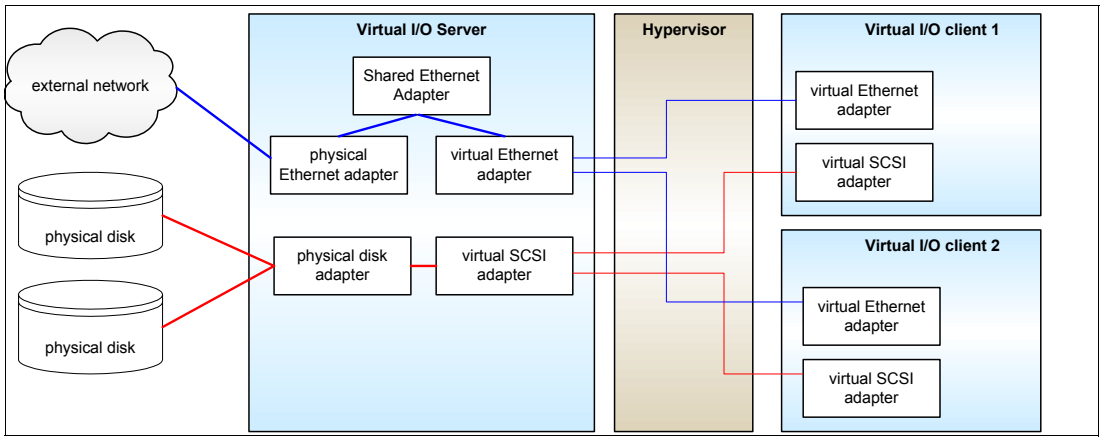


Figure 3-2 Architectural view of the Virtual I/O Server

Because the Virtual I/O server is an operating system-based appliance server, redundancy for physical devices attached to the Virtual I/O Server can be provided by using capabilities such as Multipath I/O and IEEE 802.3ad Link Aggregation.

Installation of the Virtual I/O Server partition is performed from a special system backup DVD that is provided to clients that order any PowerVM edition. This dedicated software is only for the Virtual I/O Server (and IVM in case it is used) and is only supported in special Virtual I/O Server partitions. Two major functions are provided with the Virtual I/O Server: a Shared Ethernet Adapter and Virtual SCSI.

Shared Ethernet Adapter

A Shared Ethernet Adapter (SEA) can be used to connect a physical Ethernet network to a virtual Ethernet network. The Shared Ethernet Adapter provides this access by connecting the internal Hypervisor VLANs with the VLANs on the external switches. Because the Shared Ethernet Adapter processes packets at layer 2, the original MAC address and VLAN tags of the packet are visible to other systems on the physical network. IEEE 802.1 VLAN tagging is supported.

The Shared Ethernet Adapter also provides the ability for several client partitions to share one physical adapter. Using an SEA, you can connect internal and external VLANs using a physical adapter. The Shared Ethernet Adapter service can only be hosted in the Virtual I/O Server, not in a general purpose AIX or Linux partition, and acts as a layer-2 network bridge to securely transport network traffic between virtual Ethernet networks (internal) and one or more (EtherChannel) physical network adapters (external). These virtual Ethernet network adapters are defined by the POWER Hypervisor on the Virtual I/O Server

Tip: A Linux partition can provide bridging function as well, by using the **`brctl`** command.

Figure 3-3 on page 83 shows a configuration example of an SEA with one physical and two virtual Ethernet adapters. An SEA can include up to 16 virtual Ethernet adapters on the Virtual I/O Server that share the same physical access.

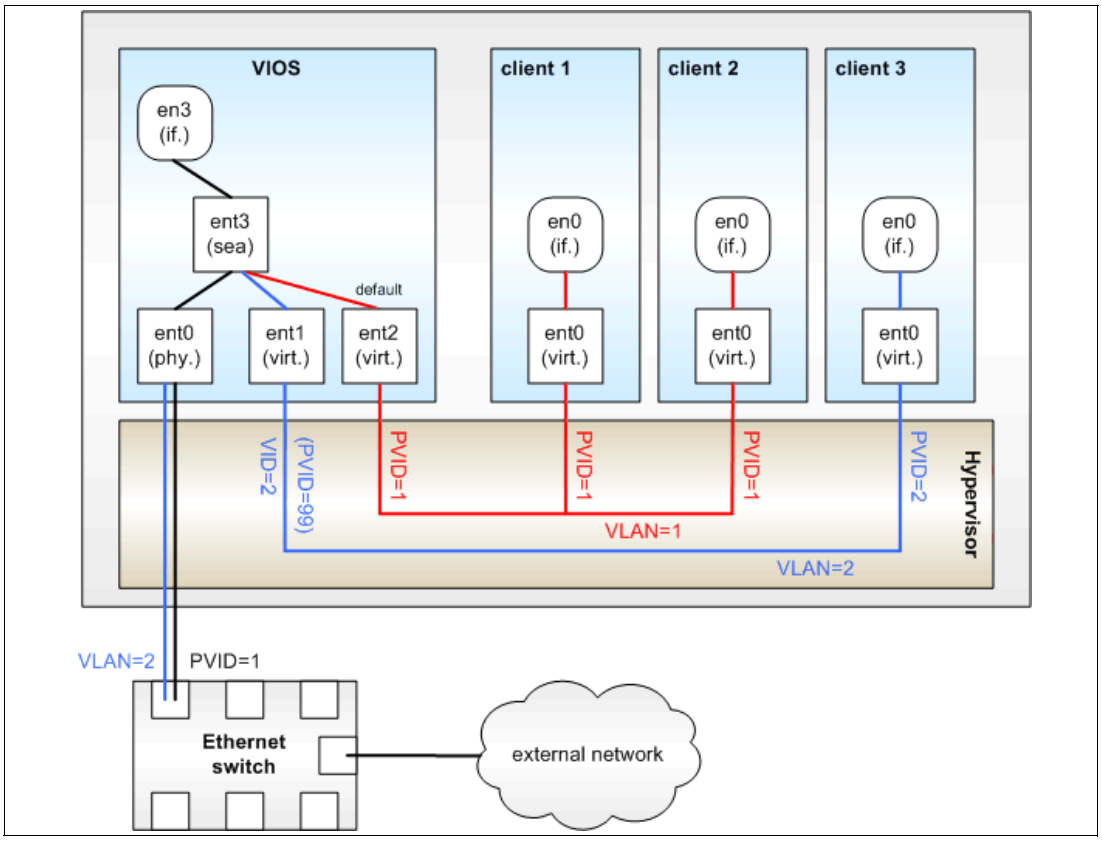


Figure 3-3 Architectural view of a Shared Ethernet Adapter

A single SEA setup can have up to 16 Virtual Ethernet trunk adapters and each virtual Ethernet trunk adapter can support up to 20 VLAN networks. Therefore, it is possible for a single physical Ethernet to be shared between 320 internal VLAN. The number of shared Ethernet adapters that can be set up in a Virtual I/O server partition is limited only by the resource availability as there are no configuration limits.

Unicast, broadcast, and multicast is supported, so protocols that rely on broadcast or multicast, such as Address Resolution Protocol (ARP), Dynamic Host Configuration Protocol (DHCP), Boot Protocol (BOOTP), and Neighbor Discovery Protocol (NDP) can work across an SEA.

Note: A Shared Ethernet Adapter does not need to have an IP address configured to be able to perform the Ethernet bridging functionality. It is very convenient to configure IP on the Virtual I/O Server. This is because the Virtual I/O Server can then be reached by TCP/IP, for example, to perform dynamic LPAR operations or to enable remote login. This can be done either by configuring an IP address directly on the SEA device, or on an additional virtual Ethernet adapter in the Virtual I/O Server. This leaves the SEA without the IP address, allowing for maintenance on the SEA without losing IP connectivity in case SEA failover is configured.

For a more detailed discussion about virtual networking, see:

http://www.ibm.com/servers/aix/whitepapers/aix_vn.pdf

Virtual SCSI

Virtual SCSI is used to refer to a virtualized implementation of the SCSI protocol. Virtual SCSI is based on a client/server relationship. The Virtual I/O Server logical partition owns the physical resources and acts as server or, in SCSI terms, target device. The client logical partitions access the virtual SCSI backing storage devices provided by the Virtual I/O Server as clients.

The virtual I/O adapters (virtual SCSI server adapter and a virtual SCSI client adapter) are configured using an HMC or through the Integrated Virtualization Manager on smaller systems. The virtual SCSI server (target) adapter is responsible for executing any SCSI commands it receives. It is owned by the Virtual I/O Server partition. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN attached devices and LUNs that are assigned to the client partition. The provisioning of virtual disk resources is provided by the Virtual I/O Server.

Physical disks presented to the Virtual/O Server can be exported and assigned to a client partition in a number of different ways:

- The entire disk is presented to the client partition
- The disk is divided into several logical volumes, these can be presented to a single client or multiple different clients
- As of Virtual I/O Server 1.5, files can be created on these disks and file backed storage devices can be created

The Logical volumes or files can be assigned to different partitions. Therefore, virtual SCSI enables sharing of adapters as well as disk devices.

Figure 3-4 shows an example where one physical disk is divided into two logical volumes by the Virtual I/O Server. Each of the two client partitions is assigned one logical volume, which is then accessed through a virtual I/O adapter (VSCSI Client Adapter). Inside the partition, the disk is seen as a normal hdisk.

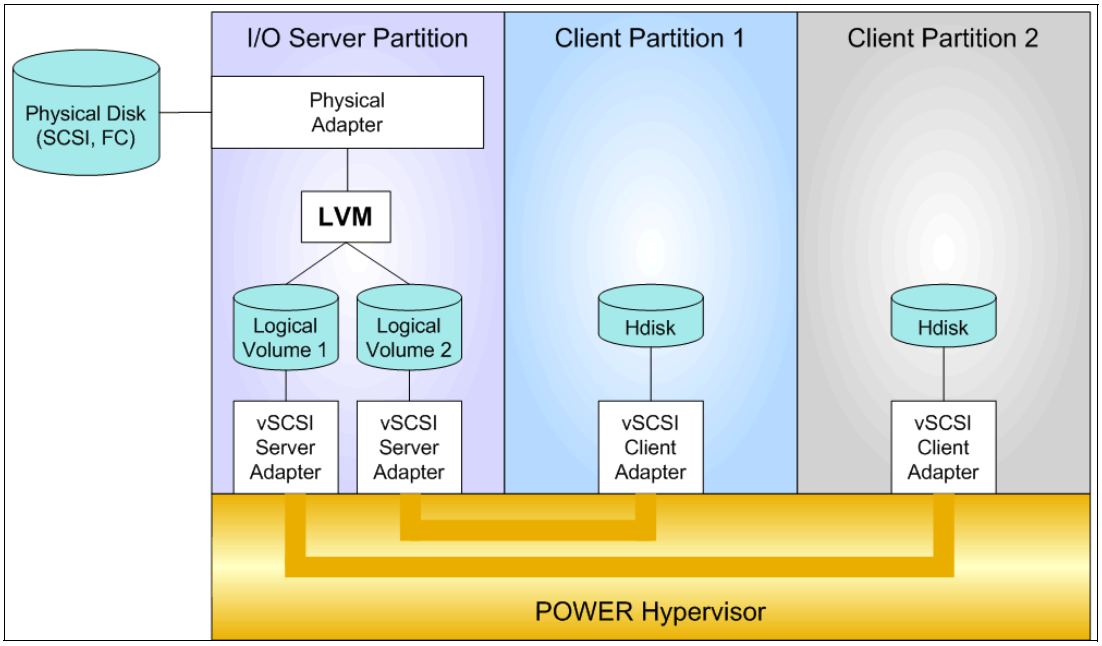


Figure 3-4 Architectural view of virtual SCSI

At the time of writing, virtual SCSI supports Fibre Channel, parallel SCSI, iSCSI, SAS, SCSI RAID devices and optical devices, including DVD-RAM and DVD-ROM. Other protocols such as SSA and tape devices are not supported.

For more information about the specific storage devices supported for Virtual I/O Server, see:

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html

Virtual I/O Server function

Virtual I/O Server includes a number of features, including monitoring solutions:

- Support for Live Partition Mobility on POWER6 processor-based systems with the PowerVM Enterprise Edition. More information about Live Partition Mobility can be found on 3.3.4, “PowerVM Live Partition Mobility” on page 87.
- Support for virtual SCSI devices backed by a file. These are then accessed as standard SCSI-compliant LUNs.
- Virtual I/O Server Expansion Pack with additional security functions like Kerberos (Network Authentication Service for users and Client and Server Applications), SNMP v3 (Simple Network Management Protocol) and LDAP (Lightweight Directory Access Protocol client functionality).
- System Planning Tool (SPT) and Workload Estimator are designed to ease the deployment of a virtualized infrastructure. More on the System Planning Tool in section 3.4, “System Planning Tool” on page 91.
- IBM Systems Director and a number of preinstalled Tivoli® agents are included like Tivoli Identity Manager (TIM) in order to allow easy integration into an existing Tivoli Systems Management infrastructure, and Tivoli Application Dependency Discovery Manager (ADDM) which creates and maintains automatically application infrastructure maps including dependencies, change histories and deep configuration values.
- vSCSI eRAS
- Additional Command Line Interface (CLI) statistics in **svmon**, **vmstat**, **fcstat** and **topas**
- Monitoring solutions to help manage and monitor the Virtual I/O Server and shared resources. New commands and views provide additional metrics for memory, paging, processes, Fibre Channel HBA statistics and virtualization.

For more information on the Virtual I/O Server and its implementation, refer to *PowerVM virtualization on IBM System p, Introduction and Configuration,* SG24-7940.

3.3.3 PowerVM Lx86

The IBM PowerVM Lx86 feature creates a virtual x86 Linux application environment on POWER processor-based systems, so most 32-bit x86 Linux applications can run without requiring clients or ISVs to recompile the code. This brings new benefits to organizations who want the reliability and flexibility of consolidating (through virtualization) on Power Systems and use applications that have not yet been ported to the platform.

PowerVM Lx86 dynamically translates x86 instructions to Power Architecture instructions, operating much like the Just-in-time compiler (JIT) in a Java™ system. The technology creates an environment in which the applications being translated run on the new target platform, in this case Linux on POWER. This environment encapsulates the application and runtime libraries and runs them on the Linux on POWER operating system kernel. These applications can be run side by side with POWER native applications on a single system image and do not require a separate partition.

Figure 3-5 shows the diagram of the Linux x86 application environment.

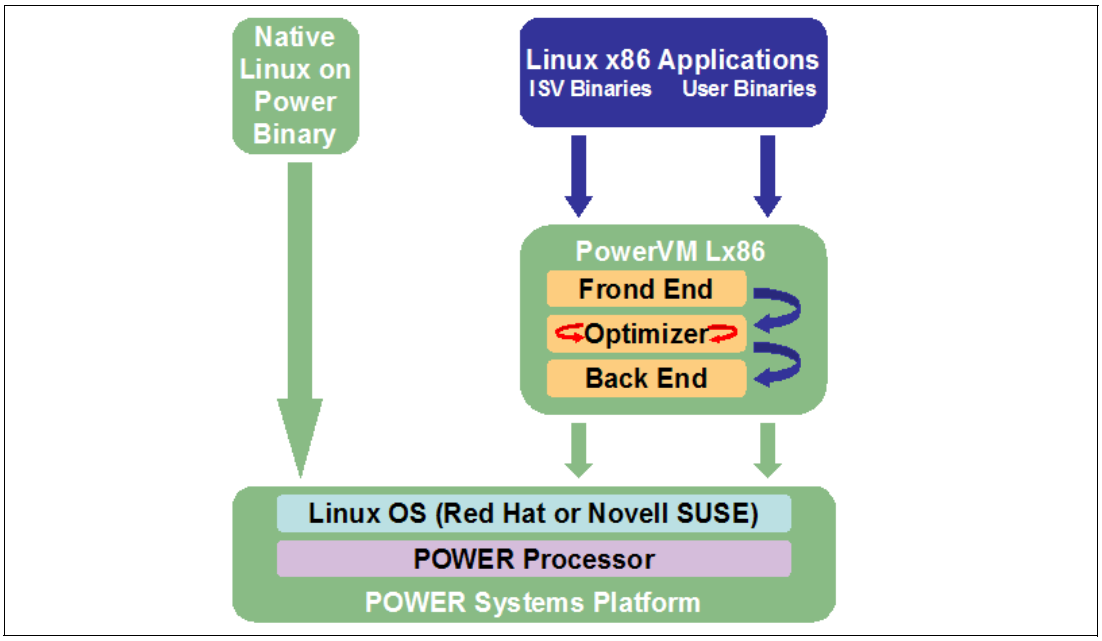


Figure 3-5 Diagram of the Linux x86 Application Environment

Supported Operating Systems

PowerVM Lx86 version 1.1 will support the following Linux on POWER operating systems:

- Red Hat Enterprise Linux 4 (RHEL 4) for POWER version 4.4 and 4.5. Also, x86 Linux applications running on RHEL 4.3 are supported.
- SUSE Linux Enterprise Server 9 (SLES 9) for POWER Service Pack 3
- SUSE Linux Enterprise Server 10 (SLES 10) for POWER Service Pack 1

Note:

- PowerVM LX86 is supported under the VIOS Software Maintenance Agreement (SWMA).
- When using PowerVM Lx86 on an IBM System p POWER6 processor-based system only SLES 10 with SP1 and RHEL 4.5 are supported.
- Make sure the x86 version is the same as your Linux on POWER version. Do not try to use any other version because it is unlikely to work. One exception is with Red Hat Enterprise Linux, both the Advanced Server and Enterprise Server option at the correct release will work.

As stated in a previous paragraph, PowerVM Lx86 runs most x86 Linux applications, but PowerVM Lx86 cannot run applications that:

- Directly access hardware devices (for example, graphics adapters)
- Require nonstandard kernel module access or use kernel modules not provided by the Linux for POWER operating system distribution
- Do not use only the Intel® IA-32 instruction set architecture as defined by the 1997 Intel Architecture Software Developer's Manual consisting of Basic Architecture (Order Number 243190), Instruction Set Reference Manual (Order Number 243191) and the System Programming Guide (Order Number 243192) dated 1997

- Do not run correctly on Red Hat Enterprise Linux 4 starting with version 4.3 or Novell SUSE Linux Enterprise Server (SLES) 9 starting with version SP3 or Novell SLES 10;
- Are x86 Linux specific system administration or configuration tools.

For more information about PowerVM Lx86 please refer to *Getting started with PowerVM Lx86*, REDP-4298.

3.3.4 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility allows you to move a running logical partition, including its operating system and running applications, from one system to another without any shutdown or without disrupting the operation of that logical partition. Inactive partition mobility allows you to move a powered off logical partition from one system to another.

Partition mobility provides systems management flexibility and improves system availability:

- Avoid planned outages for hardware or firmware maintenance by moving logical partitions to another server and then performing the maintenance. Live partition mobility can help lead to zero downtime maintenance because you can use it to work around scheduled maintenance activities.
- Avoid downtime for a server upgrade by moving logical partitions to another server and then performing the upgrade. This allows your end users to continue their work without disruption.
- Preventive failure management: If a server indicates a potential failure, you can move its logical partitions to another server before the failure occurs. Partition mobility can help avoid unplanned downtime.
- Server optimization:
  - You can consolidate workloads running on several small, under-used servers onto a single large server.
  - Deconsolidation: You can move workloads from server to server to optimize resource use and workload performance within your computing environment. With active partition mobility, you can manage workloads with minimal downtime.

Mobile partition's operating system requirements

The operating system running in the mobile partition has to be AIX or Linux. The Virtual I/O Server logical partition hat to be at least at the 1.5 release level. However, the Virtual I/O Server partition itself cannot be migrated. The operating system must be at one of the following levels:

- AIX 5L V5.3 with 5300-07 Technology Level or later
- AIX V6.1 or later
- Red Hat Enterprise Linux Version V5.1 or later
- SUSE Linux Enterprise Services 10 (SLES 10) Service Pack 1 or later

Previous versions of AIX and Linux can participate in inactive partition mobility, if the operating systems support virtual devices and IBM System p POWER6 processor-based systems.

Source and destination system requirements

The source partition must be one that only has virtual devices. If there are any physical devices in its allocation, they must be removed before the validation or migration is initiated.

The hypervisor must support the Partition Mobility functionality also called migration process. POWER 6 processor-based hypervisors have this capability and firmware must be at firmware level eFW3.2 or later. Source and destination systems could have different firmware levels, but they must be compatible with each other.

The virtual I/O server on the source system provides the access to the clients resources and must be identified as a Mover Service Partition (MSP). The VASI device , Virtual Asynchronous Services Interface (VASI) allows the mover service partition to communicate with the hypervisor; it is created and managed automatically by the HMC and will be configured on both the source and destination Virtual I/O Servers designated as the mover service partitions for the mobile partition to participate in active mobility. Other requirements include a similar Time of Day on each server, systems shouldn't be running on battery power, shared storage (external hdisk with reserve_policy=no_reserve), and all logical partitions should be on the same open network with RMC established to the HMC.

The HMC is used to configure, validate and to orchestrate. You will use the HMC to configure the Virtual I/O Server as an MSP and to configure the VASI device. An HMC wizard validates your configuration and identifies things which will cause the migration to fail. During the migration, the HMC controls all phases of the process.

For more information about Live Partition Mobility and how to implement it, refer to *IBM System p Live Partition Mobility*, SG24-7460.

3.3.5 PowerVM AIX 6 Workload Partitions

Workload partitions will provide a way for clients to run multiple applications inside the same instance of an AIX operating system while providing security and administrative isolation between applications. Workload partitions complement logical partitions and can be used in conjunction with logical partitions and other virtualization mechanisms, if desired. Workload partitions (WPAR) is a software-base virtualization capability of AIX V6 that can improve administrative efficiency by reducing the number of AIX operating system instances that must be maintained and can increase the overall utilization of systems by consolidating multiple workloads on a single system and is designed to improve cost of ownership.

The use of workload partitions is optional, therefore programs will run as before if run in the Global environment (AIX instance). This Global environment owns all the physical resources (like adapters, memory, disks, processors) of the logical partition.

Note Workload partitions are only supported with AIX V6.

Workload partitions are separate regions of application space. and therefore they allow users to create multiple software-based partitions on top of a single AIX instance. This approach enables high levels of flexibility and capacity utilization for applications executing heterogeneous workloads, and simplifies patching and other operating system maintenance tasks.

There are two types of workload partitions:

- System workload partitions - these are autonomous virtual system environments with their own private root file systems, users and groups, login, network space and administrative domain. It represents a partition within the operating system isolating runtime resources such as memory, CPU, user information, or file system to specific application processes. Each System workload partition has its own unique set of users, groups and network addresses. It is integrated with he Role Based Access control (RBAC). Inter-process communication for a process in a workload partition is restricted to those processes in the

same workload partition. The systems administrator accesses the workload partition via the administrator console or via regular network tools such as **telnet** or **ssh**. The system workload partition is removed only when requested.

- Application workload partitions - these are light weight workload partitions because there is no system services involved, there is no file system isolation since it uses the global environment system file system, Telnet is not supported but access through console login is available. Once the application process or processes are finished the workload partition is stopped.

For a detailed list of the workload partitions concepts and function, refer to *Introduction to Workload Partition Management in IBM AIX Version 6*, SG24-7431.

3.3.6 PowerVM AIX 6 Workload Partition Manager

IBM PowerVM AIX 6 Workload Partition Manager (WPAR Manager) is a platform management solution that provides a centralized point of control for managing workload partitions (WPARs) across a collection of managed systems running AIX. It is an optional product designed to facilitate the management of WPARs as well as provide advanced features such as policy based application mobility for automation of workload partitions relocation based on current performance state.

The workload partition manager is an intuitive graphical user interface based tool designed to provide a centralized interface for administration of WPAR instances across multiple systems. By deploying the workload partitions manager, users are able to take full advantage of workload partitions technology by leveraging the following features:

- Basic life cycle management

  Create, start, stop, and delete WPAR instances

- Manual WPAR mobility

  User initiated relocation of WPAR instances

- Creation and administration of mobility policies

  User defined policies governing automated relocation of WPAR instances based on performance state

- Creation of compatibility criteria on a per WPAR basis

  User defined criteria based on compatibility test results gathered by the WPAR Manager

- Administration of migration domains

  Creation and management of server groups associated to specific WPAR instances which establish which servers would be appropriate as relocation targets

- Server profile ranking

  User defined rankings of servers for WPAR relocation based on performance state

- Reports based on historical performance

  Performance metrics gathered by WPAR Manager for both servers and WPAR instances

- Event logs and error reporting

  Detailed information related to actions taken during WPAR relocation events and other system operations

- Inventory and automated discovery

  Complete inventory of WPAR instances deployed on all servers with WPAR Manager Agents installed whether created by the WPAR Manager or through the command line interface (CLI) on the local system console

Note: The IBM PowerVM Workload Partition Manager for AIX 6 is a separate, optional product as part of the IBM PowerVM suite (5765-WPM).

3.3.7 Operating System support for PowerVM

Table 3-3 lists AIX 5L, AIX V6.1 and Linux support for PowerVM.

Table 3-3 PowerVM features supported by AIX and Linux

Feature	AIX V5.3	AIX V6.1	RHEL V4.5 for POWER	RHEL V5.1 for POWER	SLES V10 SP1 for POWER
DLPAR operations [1]	Y	Y	Y	Y	Y
Capacity Upgrade on Demand [2]	Y	Y	Y	Y	Y
Micro-Partitioning	Y	Y	Y	Y	Y
Shared Dedicated Capacity	Y	Y	N	Y	Y
Multiple Shared Processor Pools	Y	Y	N	Y	Y
Virtual I/O Server	Y	Y	Y	Y	Y
IVM	Y	Y	Y	Y	Y
Virtual SCSI	Y	Y	Y	Y	Y
Virtual Ethernet	Y	Y	Y	Y	Y
Live Partition Mobility	Y	Y	N	Y	Y
Workload Partitions	N	Y	N	N	N

1 Dynamic memory removal is not supported by Linux at the time of writing.
2 On selected models available.

3.4 System Planning Tool

The IBM System Planning Tool (SPT) helps you design a system or systems to be partitioned with logical partitions. You can also plan for and design non-partitioned systems using the SPT. The resulting output of your design is called a *system plan*, which is stored in a .sysplan file. This file can contain plans for a single system or multiple systems. The .sysplan file can be used:

- To create reports
- As input to the IBM configuration tool (eConfig)
- To create and deploy partitions on your system(s) automatically

The SPT is the next generation of the IBM LPAR Validation Tool (LVT). It contains all the functions from the LVT, as well as significant functional enhancements, and is integrated with the IBM Systems Workload Estimator (WLE). System plans generated by the SPT can be deployed on the system by the Hardware Management Console (HMC) or the Integrated Virtualization Manager (IVM).

Note: Ask your IBM Representative or Business Partner to use the Customer Specified Placement manufacturing option if you want to automatically deploy your partitioning environment on an new machine. SPT looks for the resource's allocation to be the same as the specified on your .sysplan file.

You can create an entirely new system configuration, or you can create a system configuration based upon any of the following:

- Performance data from an existing system that the new system is to replace
- Performance estimates that anticipates future workloads that you must support
- Sample systems that you can customize to fit your needs

Integration between the SPT and both the Workload Estimator (WLE) and IBM Performance Management (PM) allows you to create a system that is based upon performance and capacity data from an existing system or that is based on new workloads that you specify.

You can use the SPT before you order a system to determine what you must order to support your workload. You can also use the SPT to determine how you can partition a system that you already have.

The SPT is a PC-based browser application designed to be run in a standalone environment. The SPT can be used to plan solutions for the following IBM systems:

- IBM Power Systems
- System p5 and System i5™
- eServer p5 and eServer i5
- OpenPower®
- iSeries® 8xx and 270 models

We recommend that you use the IBM System Planning Tool to estimate POWER Hypervisor requirements and determine the memory resources that are required for all partitioned and non-partitioned servers.

Figure 3-6 on page 92 shows the estimated Hypervisor memory requirements based on sample partition requirements.

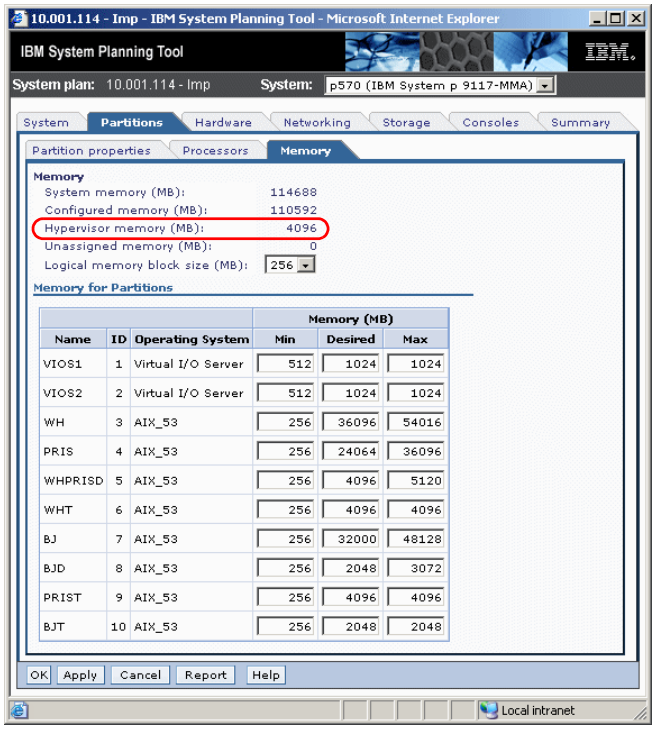


Figure 3-6 IBM System Planning Tool window showing Hypervisor memory requirements

Note: In previous releases of the SPT, you could view an HMC or IVM system plan, but you could not edit the plan. The SPT now allows you to convert an HMC or IVM system plan into a format that will allow you to edit the plan in the SPT.

Also note that SPT 2.0 will be the last release that will support .lvt and .xml files. Users should load their old .lvt and .xml plans and save them as .sysplan files.

It is recommended that you take action prior to 03/31/2008.

The SPT and its supporting documentation can be found on the IBM System Planning Tool site at:

http://www.ibm.com/systems/support/tools/systemplanningtool/

4

Continuous availability and manageability

This chapter provides information about IBM Power Systems design features that help lower the total cost of ownership (TCO). The advanced IBM RAS (Reliability, Availability, and Service ability) technology allows the possibility to improve your architecture's TCO by reducing unplanned down time.

IBM POWER6 processor-based systems have a number of new features that enable systems to dynamically adjust when issues arise that threaten availability. Most notably, POWER6 processor-based systems introduce the POWER6 Processor Instruction Retry suite of tools, which includes Processor Instruction Retry, Alternate Processor Recovery, Partition Availability Prioritization, and Single Processor Checkstop. Taken together, in many failure scenarios these features allow a POWER6 processor-based system to recover transparently without an impact on a partition using a failing core.

This chapter includes several features that are based on the benefits available when using AIX as the operating system. Support of these features when using Linux can vary.

4.1 Reliability

Highly reliable systems are built with highly reliable components. On IBM POWER6 processor-based systems, this basic principle is expanded upon with a clear design for reliability architecture and methodology. A concentrated, systematic, architecture-based approach is designed to improve overall system reliability with each successive generation of system offerings.

4.1.1 Designed for reliability

Systems designed with fewer components and interconnects have fewer opportunities to fail. Simple design choices such as integrating two processor cores on a single POWER chip can dramatically reduce the opportunity for system failures. In this case, a 16-core server will include half as many processor chips (and chip socket interfaces) as with a

© Copyright IBM Corp. 2008. All rights reserved.

single-CPU-per-processor design. Not only does this reduce the total number of system components, it reduces the total amount of heat generated in the design, resulting in an additional reduction in required power and cooling components.

Parts selection also plays a critical role in overall system reliability. IBM uses three grades of components, with grade 3 defined as industry standard (off-the-shelf). As shown in Figure 4-1, using stringent design criteria and an extensive testing program, the IBM manufacturing team can produce grade 1 components that are expected to be 10 times more reliable than industry standard. Engineers select grade 1 parts for the most critical system components. Newly introduced organic packaging technologies, rated grade 5, achieve the same reliability as grade 1 parts.

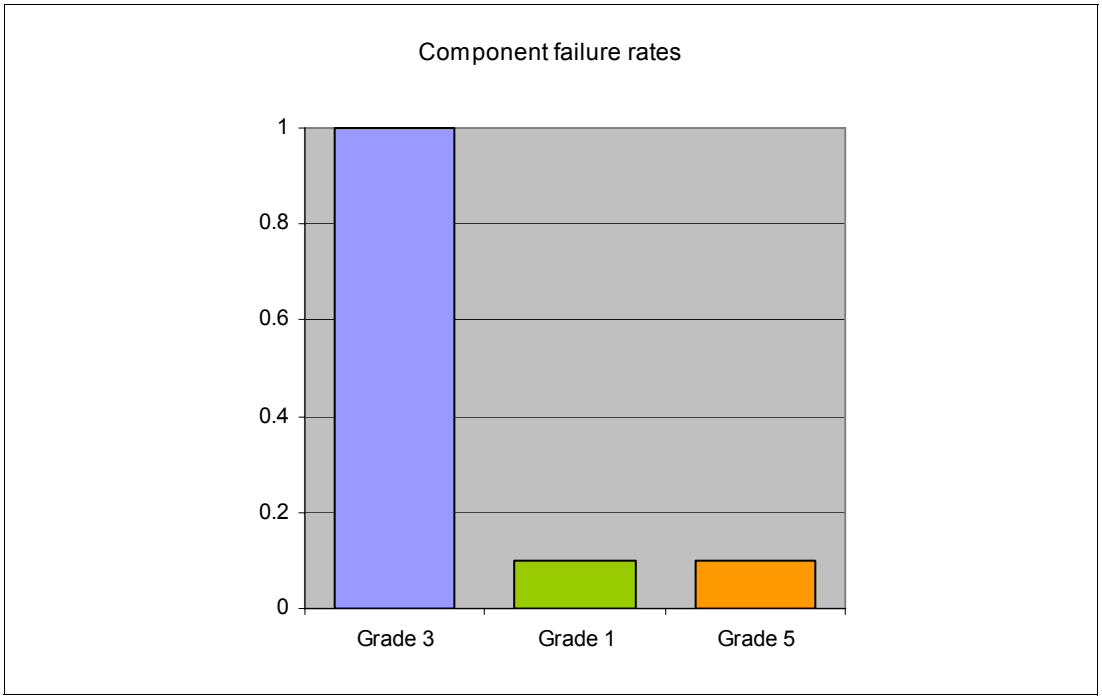


Figure 4-1 Component failure rates

4.1.2 Placement of components

Packaging is designed to deliver both high performance and high reliability. For example, the reliability of electronic components is directly related to their thermal environment, that is, large decreases in component reliability are directly correlated with relatively small increases in temperature, POWER6 processor-based systems are carefully packaged to ensure adequate cooling. Critical system components such as the POWER6 processor chips are positioned on printed circuit cards so they receive fresh air during operation. In addition, POWER6 processor-based systems are built with redundant, variable-speed fans that can automatically increase output to compensate for increased heat in the central electronic complex.

4.1.3 Redundant components and concurrent repair

High-opportunity components, or those that most affect system availability, are protected with redundancy and the ability to be repaired concurrently.

The use of redundant part allows the system to remain operational:

- Redundant service processor

  Redundant service processor function for managing service processors when one fails, is available for system configurations with two or more CEC enclosures. Redundant Service Processor function requires that the HMC be attached to the Service Interface Card in both CEC enclosure 1 and CEC enclosure 2. The Service Interface Card in these two enclosures must be connected using an external Power Control cable (FC 6006 or similar).

- Processor power regulators

  A third Processor Power Regulator is required to provide redundant power support to either one or two processor cards in the enclosure. All CEC enclosures are shipped with three Processor Power Regulators (FC 5625) except for the system configurations with one or two FC 5620 processors in a single CEC enclosure.

- Redundant spare memory bits in cache, directories and main memory
- Redundant and hot-swap cooling
- Redundant and hot-swap power supplies

  For maximum availability it is highly recommended to connect power cords from the same system to two separate PDUs in the rack. And to connect each PDU to independent power sources.

4.1.4 Continuous field monitoring

Aided by the IBM First Failure Data Capture (FFDC) methodology and the associated error reporting strategy, commodity managers build an accurate profile of the types of failures that might occur, and initiate programs to enable corrective actions. The IBM support team also continually analyzes critical system faults, testing to determine if system firmware and maintenance procedures and tools are effectively handling and recording faults as designed. See section 4.3.1, “Detecting errors” on page 105.

4.2 Availability

IBMs extensive system of FFDC error checkers also supports a strategy of Predictive Failure Analysis®: the ability to track intermittent correctable errors and to vary components off-line before they reach the point of hard failure causing a crash.

This methodology supports IBMs autonomic computing initiative. The primary RAS design goal of any POWER processor-based server is to prevent unexpected application loss due to unscheduled server hardware outages. To accomplish this goal this system have a quality design that includes critical attributes for:

- Self-diagnose and self-correct during run time
- Automatically reconfigure to mitigate potential problems from suspect hardware
- The ability to self-heal or to automatically substitute good components for failing components

4.2.1 Detecting and deallocating failing components

Runtime correctable or recoverable errors are monitored to determine if there is a pattern of errors. If these components reach a predefined error limit, the service processor initiates an

action to deconfigure the faulty hardware, helping to avoid a potential system outage and to enhance system availability.

Persistent deallocation

To enhance system availability, a component that is identified for deallocation or deconfiguration on a POWER6 processor-based system is flagged for persistent deallocation. Component removal can occur either dynamically (while the system is running) or at boot-time (IPL), depending both on the type of fault and when the fault is detected.

In addition, runtime unrecoverable hardware faults can be deconfigured from the system after the first occurrence. The system can be rebooted immediately after failure and resume operation on the remaining stable hardware. This prevents the same faulty hardware from affecting system operation again, while the repair action is deferred to a more convenient, less critical time.

Persistent deallocation functions include:

- Processor
- Memory
- Deconfigure or bypass failing I/O adapters
- L2, L3 cache

Note: The auto-restart (reboot) option has to be enabled from the Advanced System Manager Interface or from the Operator Panel.

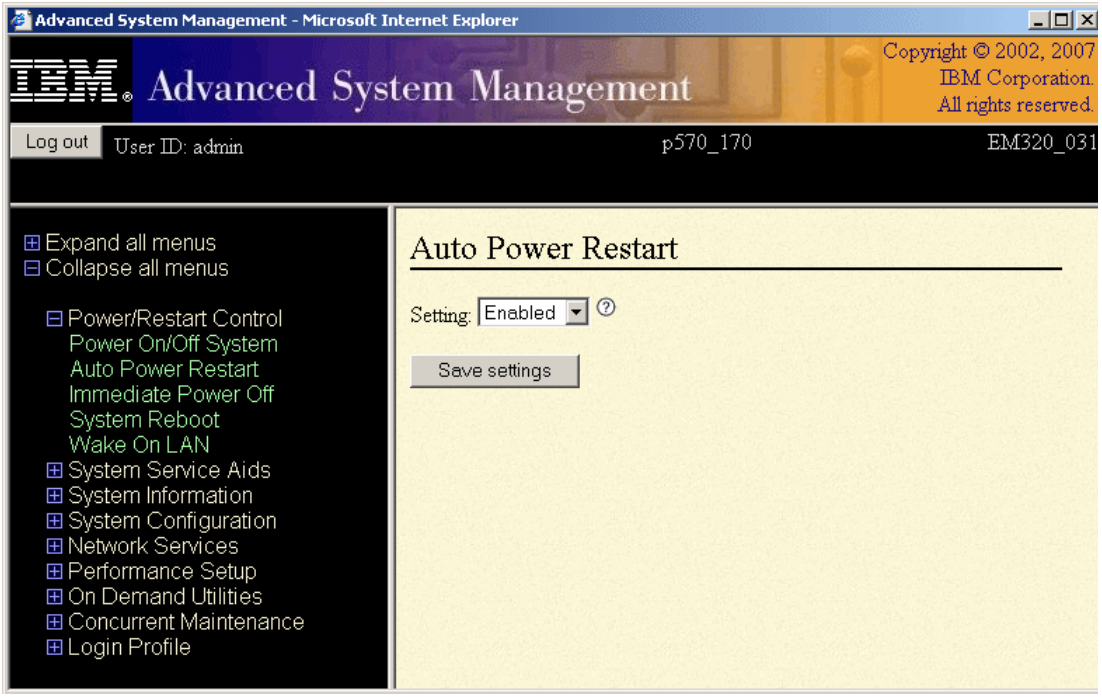


Figure 4-2 ASMI Auto Power Restart setting

Dynamic processor deallocation

Dynamic processor deallocation enables automatic deconfiguration of processor cores when patterns of recoverable errors, for example correctable errors on processor caches, are detected. Dynamic processor deallocation prevents a recoverable error from escalating to an

unrecoverable system error, which might otherwise result in an unscheduled server outage. Dynamic processor deallocation relies upon the service processor's ability to use FFDC-generated recoverable error information to notify the POWER Hypervisor when a processor core reaches its predefined error limit. Then the POWER Hypervisor, in conjunction with the operating system, redistributes the work to the remaining processor cores, deallocates the offending processor core, and continues normal operation. Even reverts from simultaneous multiprocessing to uniprocessor processing.

IBMs logical partitioning strategy allows any processor core to be shared with any partition on the system, thus enables the following sequential scenarios for processor deallocation and sparing.

1. An unlicensed Capacity on Demand (CoD) processor core is by default automatically used for Dynamic Processor Sparing
2. If no CoD processor core is available, the POWER Hypervisor attempts to locate an un-allocated core somewhere in the system
3. If no spare processor core is available, the POWER Hypervisor attempts to locate a total of 1.00 spare processing units from a shared processor pool and redistributes the workload.
4. If the requisite spare capacity is not available, the POWER Hypervisor will determine how many processing units each partition will need to relinquish to create at least 1.00 processing unit shared processor pool.
5. Once a full core equivalent is attained, the CPU deallocation event occurs.

Figure 4-3 shows a scenario where CoD processor cores are available for dynamic processor sparing.

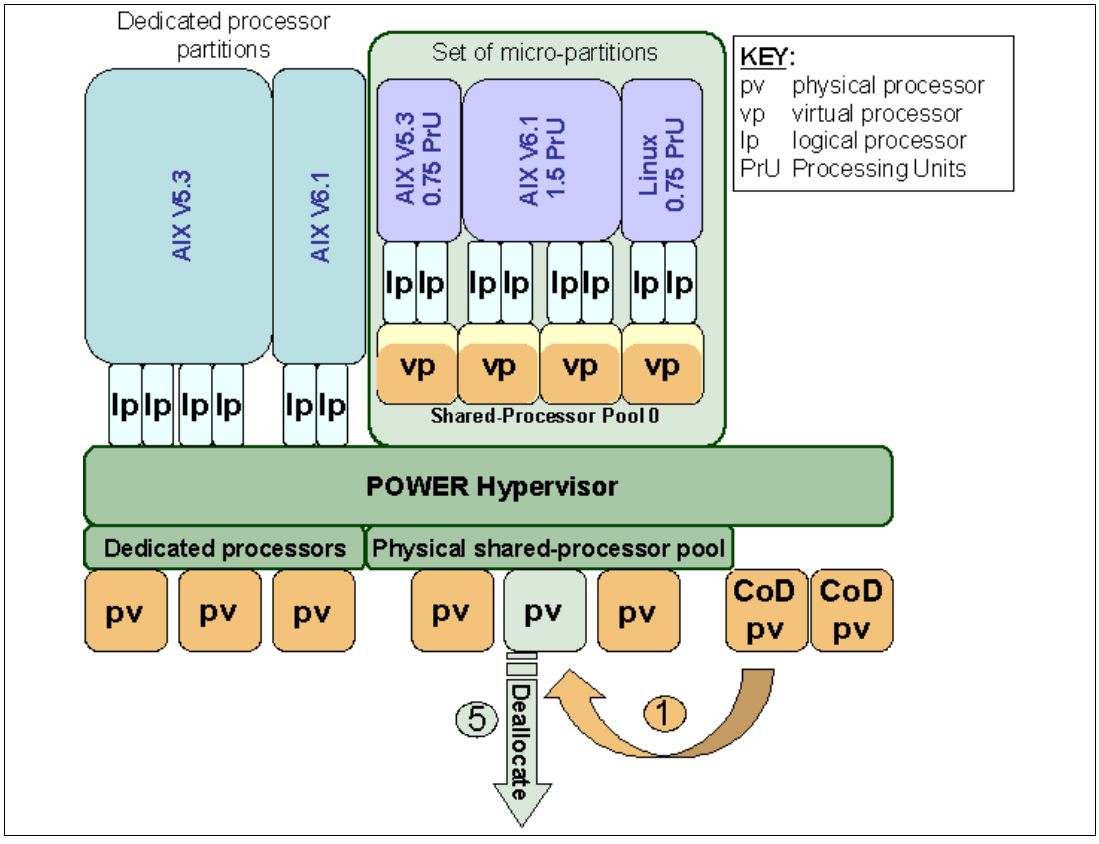


Figure 4-3 Dynamic processor deallocation and dynamic processor sparing

The deallocation event is not successful if the POWER Hypervisor and OS cannot create a full core equivalent. This result in an error message and the requirement for a system administrator to take corrective action. In all cases, a log entry is made for each partition that could use the physical core in question.

POWER6 processor instruction retry

POWER6 processor-based systems include a suite of mainframe-inspired processor instruction retry features that can significantly reduce situations that could result in checkstop.

Processor instruction retry — Automatically retry a failed instruction and continue with the task.

Alternate processor recovery — Interrupt a repeatedly failing instruction and move it to a new processor and continue with the task.

Partition availability priority — Starting with POWER6 technology, partitions receive an integer rating with the lowest priority partition rated at “0” and the highest priority partition valued at “255.” The default value is set at “127” for standard partitions and “192” for Virtual I/O Server (VIOS) partitions. Partition Availability Priorities are set for both dedicated and shared partitions. To initiate Alternate Processor Recovery when a spare processor is not available, the POWER Hypervisor uses the Partition availability priority to identify low priority partitions and keep high priority partitions running at full capacity.

Processor contained checkstop — When all the above mechanisms fail in almost all cases (excepting the POWER Hypervisor) a termination will be contained to the single partition using the failing processor core.

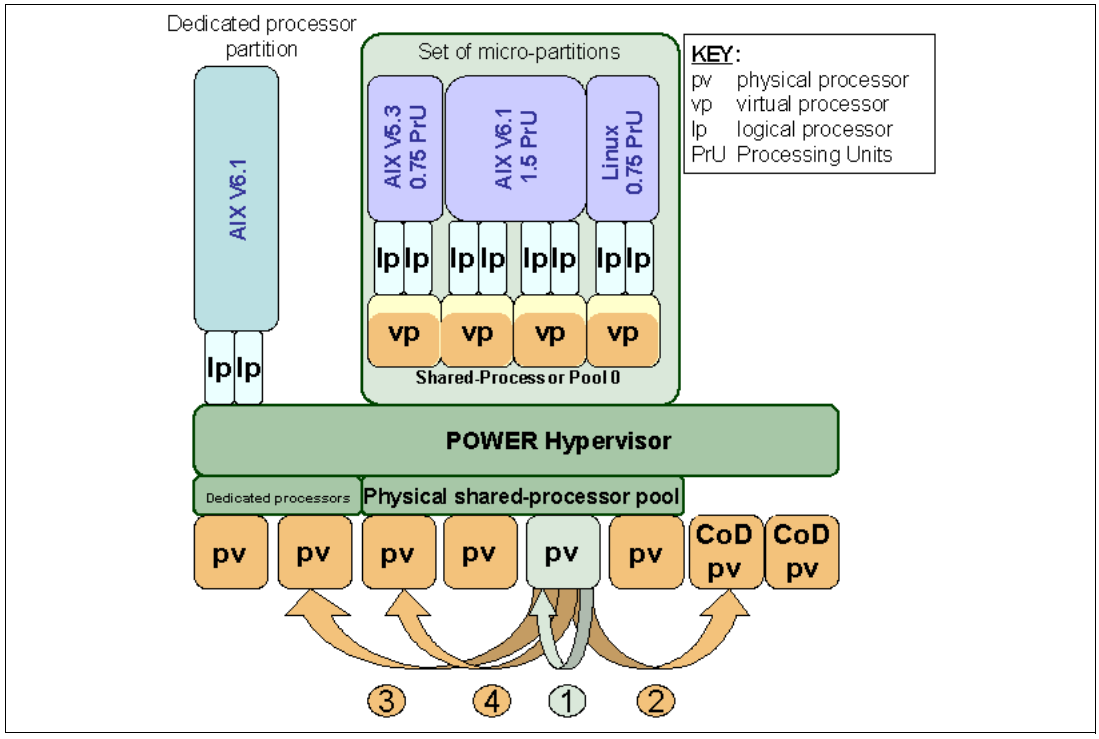


Figure 4-4 Processor instruction retry and Alternate processor recovery

Memory protection

Memory and cache arrays comprise data bit lines that feed into a memory word. A memory word is addressed by the system as a single element. Depending on the size and addressability of the memory element, each data bit line may include thousands of individual bits or memory cells. For example:

- A single memory module on a Dual Inline Memory Module (DIMM) can have a capacity of 1 Gb, and supply eight bit lines of data for an ECC word. In this case, each bit line in the ECC word holds 128 Mb behind it, corresponding to more than 128 million memory cell addresses.
- A 32 KB L1 cache with a 16-byte memory word, on the other hand, would have only 2 Kb behind each memory bit line.

A memory protection architecture that provides good error resilience for a relatively small L1 cache might be very inadequate for protecting the much larger system main store. Therefore, a variety of different protection methods are used in POWER6 processor-based systems to avoid uncorrectable errors in memory.

Memory protection plans must take into account many factors, including:

- Size
- Desired performance
- Memory array manufacturing characteristics.

POWER6 processor-based systems have a number of protection schemes designed to prevent, protect, or limit the effect of errors in main memory. These capabilities include:

Hardware scrubbing Hardware scrubbing is a method used to deal with soft errors. IBM POWER6 processor-based systems periodically address all memory locations and any memory locations with an ECC error are rewritten with the correct data.

Error correcting code Error correcting code (ECC) allows a system to detect up to two errors in a memory word and correct one of them. However, without additional correction techniques if more than one bit is corrupted, a system will fail.

Chipkill™ Chipkill is an enhancement to ECC that enables a system to sustain the failure of an entire DRAM. Chipkill spreads the bit lines from a DRAM over multiple ECC words, so that a catastrophic DRAM failure would affect at most one bit in each word. Barring a future single bit error, the system can continue indefinitely in this state with no performance degradation until the failed DIMM can be replaced.

Redundant bit steering IBM systems use redundant bit steering to avoid situations where multiple single-bit errors align to create a multi-bit error. In the event that an IBM POWER6 processor-based system detects an abnormal number of errors on a bit line, it can dynamically steer the data stored at this bit line into one of a number of spare lines.

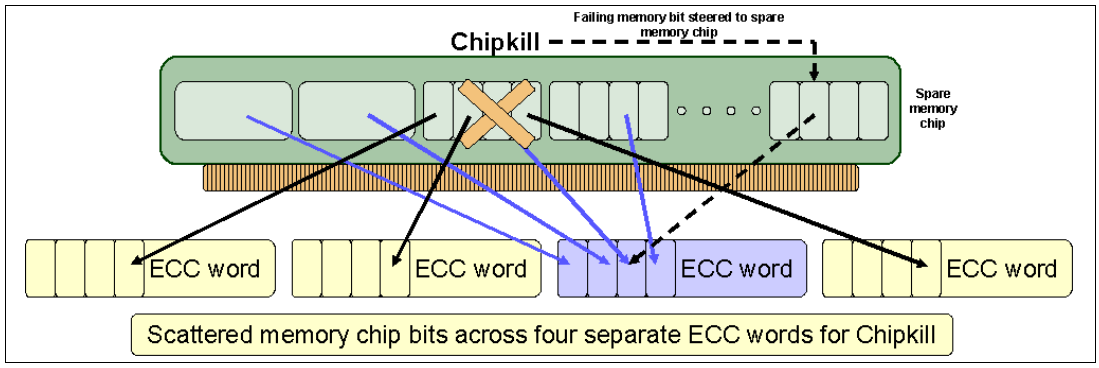


Figure 4-5 Memory protection capabilities in action

Memory page deallocation

While coincident single cell errors in separate memory chips is a statistic rarity, POWER6 processor-based systems can contain these errors using a memory page deallocation scheme for partitions running AIX and for memory pages owned by the POWER Hypervisor. If a memory address experiences an uncorrectable or repeated correctable single cell error, the Service Processor sends the memory page address to the POWER Hypervisor to be marked for deallocation.

The operating system performs memory page deallocation without any user intervention and is transparent to end users and applications.

The POWER Hypervisor maintains a list of pages marked for deallocation during the current platform IPL. During a partition IPL, the partition receives a list of all the bad pages in its address space.

In addition, if memory is dynamically added to a partition (through a Dynamic LPAR operation), the POWER Hypervisor warns the operating system if memory pages are included which need to be deallocated.

Finally, should an uncorrectable error occur, the system can deallocate the memory group associated with the error on all subsequent system reboots until the memory is repaired. This is intended to guard against future uncorrectable errors while waiting for parts replacement.

Note: Memory page deallocation handles single cell failures, but, because of the sheer size of data in a data bit line, it may be inadequate for dealing with more catastrophic failures. Redundant bit steering will continue to be the preferred method for dealing with these types of problems.

Memory control hierarchy

A memory controller on a POWER6 processor-based system is designed with four ports. Each port connects up to three DIMMs using a daisy-chained bus. The memory bus supports ECC checking on data, addresses, and command information. A spare line on the bus is also available for repair using a self-healing strategy. In addition, ECC checking on addresses and commands is extended to DIMMs on DRAMs. Because it uses a daisychained memory access topology, this system can deconfigure a DIMM that encounters a DRAM fault, without deconfiguring the bus controller, even if the bus controller is contained on the DIMM.

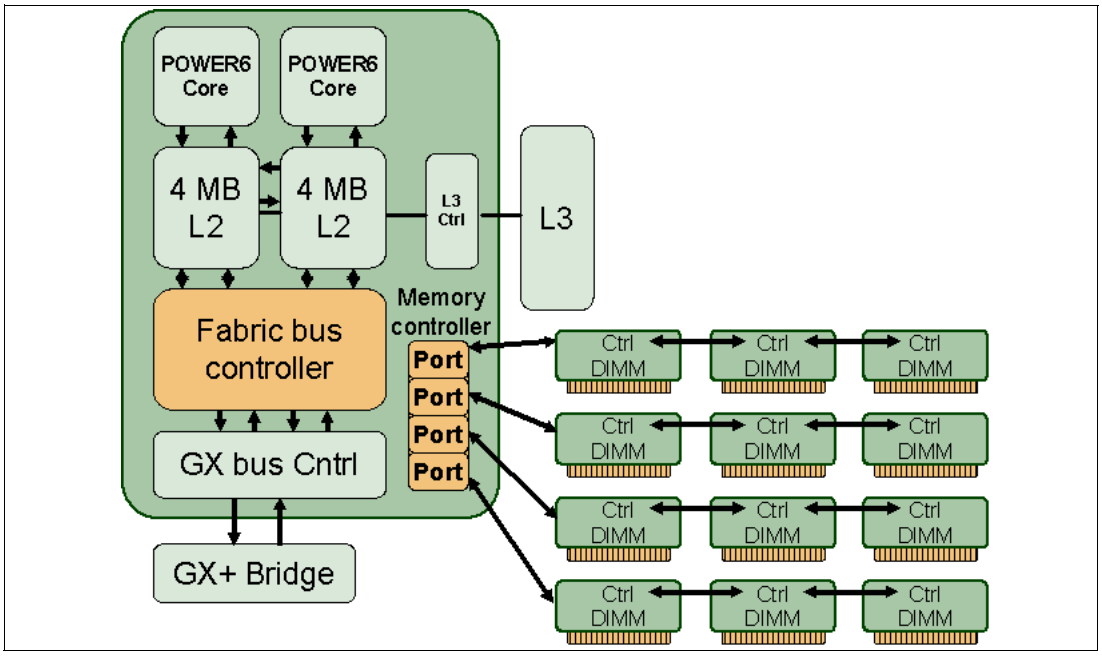


Figure 4-6 Memory control hierarchy

Memory deconfiguration

Defective memory discovered at boot time is automatically switched off, unless it is already the minimum amount required to boot. If the Service Processor detects a memory fault at boot time, it marks the affected memory as bad so it is not to be used on subsequent reboots (Memory Persistent Deallocation).

If the Service Processor identifies faulty memory in a server that includes CoD memory, the POWER Hypervisor attempts to replace the faulty memory with available CoD memory. Faulty resources are marked as deallocated and working resources are included in the active memory space. Since these activities reduce the amount of CoD memory available for future use, repair of the faulty memory should be scheduled as soon as is convenient.

Upon reboot, if not enough memory is available to meet minimum partition requirements the POWER Hypervisor will reduce the capacity of one or more partitions. The HMC receives notification of the failed component, triggering a service call.

4.2.2 Special uncorrectable error handling

While it is rare an uncorrectable data error can occur in memory or a cache. POWER6 processor-based systems attempt to limit, to the least possible disruption, the impact of an uncorrectable error using a well-defined strategy that first considers the data source. Sometimes, an uncorrectable error is temporary in nature and occurs in data that can be recovered from another repository. For example:

- Data in the instruction L1 cache is never modified within the cache itself. Therefore, an uncorrectable error discovered in the cache is treated like an ordinary cache miss, and correct data is loaded from the L2 cache.
- The L3 cache of the POWER6 processor-based systems can hold an unmodified copy of data in a portion of main memory. In this case, an uncorrectable error would simply trigger a reload of a cache line from main memory.

In cases where the data cannot be recovered from another source, a technique called Special Uncorrectable Error (SUE) handling is used to determine whether the corruption is truly a threat to the system. If, as may sometimes be the case, the data is never actually used but is simply over-written, then the error condition can safely be voided and the system will continue to operate normally.

When an uncorrectable error is detected, the system modifies the associated ECC word, thereby signaling to the rest of the system that the “standard” ECC is no longer valid. The Service Processor is then notified, and takes appropriate actions. When running AIX V5.2 or greater or Linux[1] and a process attempts to use the data, the OS is informed of the error and terminates only the specific user program.

It is only in the case where the corrupt data is used by the POWER Hypervisor that the entire system must be rebooted, thereby preserving overall system integrity.

Depending upon system configuration and source of the data, errors encountered during I/O operations may not result in a machine check. Instead, the incorrect data is handled by the processor host bridge (PHB) chip. When the PHB chip detects a problem it rejects the data, preventing data being written to the I/O device. The PHB then enters a freeze mode halting normal operations. Depending on the model and type of I/O being used, the freeze may include the entire PHB chip, or simply a single bridge. This results in the loss of all I/O operations that use the frozen hardware until a power-on reset of the PHB. The impact to partition(s) depends on how the I/O is configured for redundancy. In a server configured for fail-over availability, redundant adapters spanning multiple PHB chips could enable the system to recover transparently, without partition loss.

4.2.3 Cache protection mechanisms

POWER6 processor-based systems are designed with cache protection mechanisms, including cache line delete in both L2 and L3 arrays, Processor Instruction Retry and Alternate Processor Recovery protection on L1-I and L1-D, and redundant “Repair” bits in L1-I, L1-D, and L2 caches, as well as L2 and L3 directories.

L1 instruction and data array protection

The POWER6 processor’s instruction and data caches are protected against temporary errors using the POWER6 Processor Instruction Retry feature and against solid failures by Alternate Processor Recovery, both mentioned earlier. In addition, faults in the SLB array are recoverable by the POWER Hypervisor.

L2 Array Protection

On a POWER6 processor-based system, the L2 cache is protected by ECC, which provides single-bit error correction and double-bit error detection. Single-bit errors are corrected before forwarding to the processor, and subsequently written back to L2. Like the other data caches and main memory, uncorrectable errors are handled during run-time by the Special Uncorrectable Error handling mechanism. Correctable cache errors are logged and if the error reaches a threshold, a Dynamic Processor Deallocation event is initiated.

Starting with POWER6 processor-based systems, the L2 cache is further protected by incorporating a dynamic cache line delete algorithm similar to the feature used in the L3 cache. Up to six L2 cache lines may be automatically deleted. It is not likely that deletion of a few cache lines will adversely affect server performance. When six cache lines have been repaired, the L2 is marked for persistent deconfiguration on subsequent system reboots until it can be replaced.

[1] SLES 10 SP1 or later, and in RHEL 4.5 or later (including RHEL 5.1).

L3 Array Protection

In addition to protection through ECC and Special Uncorrectable Error handling, the L3 cache also incorporates technology to handle memory cell errors via a special cache line delete algorithm. During system run-time, a correctable error is reported as a recoverable error to the Service Processor. If an individual cache line reaches its predictive error threshold, it will be dynamically deleted. The state of L3 cache line delete will be maintained in a deallocation record, and will persist through future reboots. This ensures that cache lines varied offline by the server will remain offline should the server be rebooted, and don't need to be rediscovered each time. These faulty lines cannot then cause system operational problems.

A POWER6 processor-based system can dynamically delete up to 14 L3 cache lines. Again, it is not likely that deletion of a few cache lines will adversely affect server performance. If this total is reached, the L3 is marked for persistent deconfiguration on subsequent system reboots until repair.

While hardware scrubbing has been a feature in POWER main memory for many years, POWER6 processor-based systems introduce a hardware-assisted L3 cache memory scrubbing feature. All L3 cache memory is periodically addressed, and any address with an ECC error is rewritten with the faulty data corrected. In this way, soft errors are automatically removed from L3 cache memory, decreasing the chances of encountering multi-bit memory errors.

4.2.4 PCI Error Recovery

IBM estimates that PCI adapters can account for a significant portion of the hardware based errors on a large server. While servers that rely on boot-time diagnostics can identify failing components to be replaced by hot-swap and reconfiguration, run time errors pose a more significant problem.

PCI adapters are generally complex designs involving extensive on-board instruction processing, often on embedded microcontrollers. They tend to use industry standard grade components with less quality than other parts of the server. As a result, they may be more likely to encounter internal microcode errors, and/or many of the hardware errors described for the rest of the server.

The traditional means of handling these problems is through adapter internal error reporting and recovery techniques in combination with operating system device driver management and diagnostics. In some cases, an error in the adapter may cause transmission of bad data on the PCI bus itself, resulting in a hardware detected parity error and causing a global machine check interrupt, eventually requiring a system reboot to continue.

In 2001, IBM introduced a methodology that uses a combination of system firmware and Extended Error Handling (EEH) device drivers that allows recovery from intermittent PCI bus errors. This approach works by recovering and resetting the adapter, thereby initiating system recovery for a permanent PCI bus error. Rather than failing immediately, the faulty device is frozen and restarted, preventing a machine check. POWER6 technology extends this capability to PCIe bus errors, and includes expanded Linux support for EEH as well.

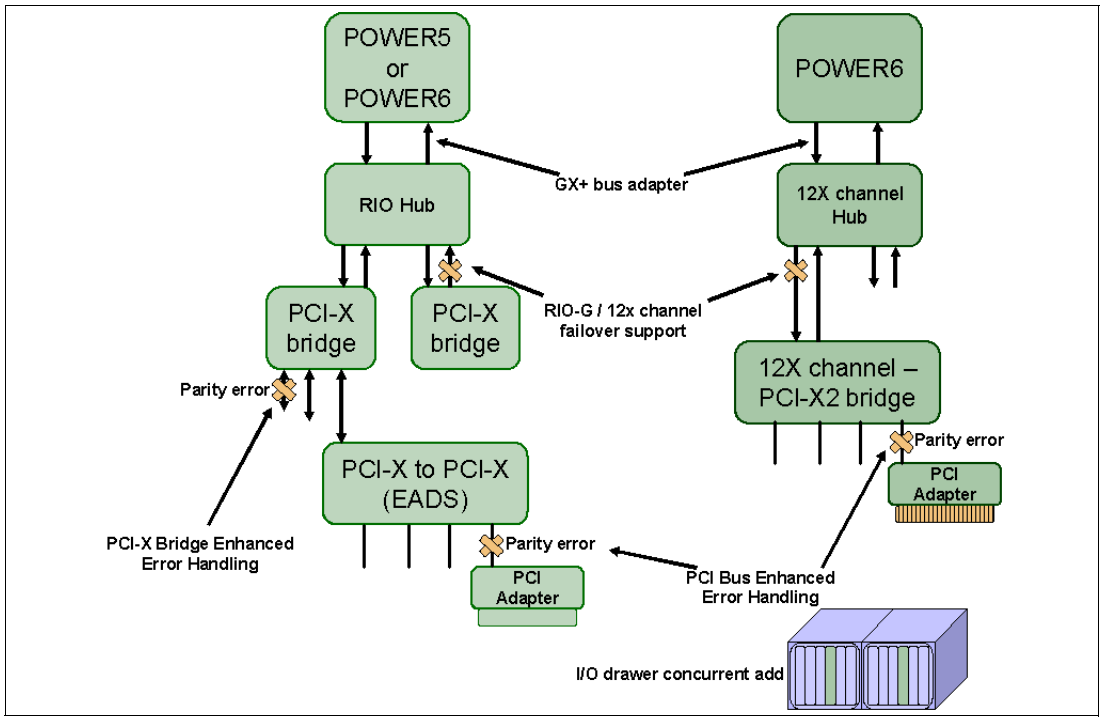


Figure 4-7 PCI error recovery

4.3 Serviceability

The IBM POWER6 Serviceability strategy evolves from, and improves upon, the service architecture deployed on the POWER5 processor-based systems. The IBM service team has enhanced the base service capabilities and continues to implement a strategy that incorporates best-of-breed service characteristics from IBMs diverse System x™, System i™, System p, and high-end System z™ offerings.

The goal of the IBM Serviceability Team is to design and provide the most efficient system service environment that incorporates:

- Easy access to service components
- On demand service education
- An automated guided repair strategy that uses common service interfaces for a converged service approach across multiple IBM server platforms

By delivering upon these goals, POWER6 processor-based systems enable faster and more accurate repair while reducing the possibility of human error.

Customer control of the service environment extends to firmware maintenance on all of the POWER6 processor-based systems, including the 570. This strategy contributes to higher systems availability with reduced maintenance costs.

This section provides an overview of the progressive steps of error detection, analysis, reporting, and repairing found in all POWER6 processor-based systems.

4.3.1 Detecting errors

The first and most crucial component of a solid serviceability strategy is the ability to accurately and effectively detect errors when they occur. While not all errors are a guaranteed threat to system availability, those that go undetected can cause problems because the system does not have the opportunity to evaluate and act if necessary. POWER6 processor-based systems employ System z server inspired error detection mechanisms that extend from processor cores and memory to power supplies and hard drives.

Service Processor

The Service Processor is a separately powered microprocessor, separate from the main instruction-processing complex. The Service Processor enables POWER Hypervisor and Hardware Management Console surveillance, selected remote power control, environmental monitoring, reset and boot features, remote maintenance and diagnostic activities, including console mirroring. On systems without a Hardware Management Console, the Service Processor can place calls to report surveillance failures with the POWER Hypervisor, critical environmental faults, and critical processing faults even when the main processing unit is inoperable. The Service Processor provides services common to modern computers such as:

- Environmental monitoring
  - The Service Processor monitors the server's built-in temperature sensors, sending instructions to the system fans to increase rotational speed when the ambient temperature is above the normal operating range.
  - Using an architected operating system interface, the Service Processor notifies the operating system of potential environmental related problems (for example, air conditioning and air circulation around the system) so that the system administrator can take appropriate corrective actions before a critical failure threshold is reached.
  - The Service Processor can also post a warning and initiate an orderly system shutdown for a variety of other conditions:
    - When the operating temperature exceeds the critical level (for example failure of air conditioning or air circulation around the system)
    - When the system fan speed is out of operational specification, for example, due to a fan failure, the system can increase speed on the redundant fans in order to compensate this failure or take other actions
    - When the server input voltages are out of operational specification.
- Mutual Surveillance
  - The Service Processor monitors the operation of the POWER Hypervisor firmware during the boot process and watches for loss of control during system operation. It also allows the POWER Hypervisor to monitor Service Processor activity. The Service Processor can take appropriate action, including calling for service, when it detects the POWER Hypervisor firmware has lost control. Likewise, the POWER Hypervisor can request a Service Processor repair action if necessary.
- Availability
  - The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced (AC power) failure.
- Fault Monitoring
  - BIST (built-in self-test) checks processor, L3 cache, memory, and associated hardware required for proper booting of the operating system, when the system is powered on at the initial install or after a hardware configuration change (e.g., an upgrade). If a

non-critical error is detected or if the error occurs in a resource that can be removed from the system configuration, the booting process is designed to proceed to completion. The errors are logged in the system nonvolatile random access memory (NVRAM). When the operating system completes booting, the information is passed from the NVRAM into the system error log where it is analyzed by error log analysis (ELA) routines. Appropriate actions are taken to report the boot time error for subsequent service if required.

One important Service Processor improvement allows the system administrator or service representative dynamic access to the Advanced Systems Management Interface (ASMI) menus. In previous generations of servers, these menus were only accessible when the system was in standby power mode. Now, the menus are available from any Web browser-enabled console attached to the Ethernet service network concurrent with normal system operation. A user with the proper access authority and credentials can now dynamically modify service defaults, interrogate Service Processor progress and error logs, set and reset guiding light LEDs, indeed, access all Service Processor functions without having to power-down the system to the standby state.

The Service Processor also manages the interfaces for connecting Uninterruptible Power Source (UPS) systems to the POWER6 processor-based systems, performing Timed Power-On (TPO) sequences, and interfacing with the power and cooling subsystem.

Error checkers

IBM POWER6 processor-based systems contain specialized hardware detection circuitry that is used to detect erroneous hardware operations. Error checking hardware ranges from parity error detection coupled with processor instruction retry and bus retry, to ECC correction on caches and system buses. All IBM hardware error checkers have distinct attributes:

- Continually monitoring system operations to detect potential calculation errors.
- Attempt to isolate physical faults based on run-time detection of each unique failure.
- Ability to initiate a wide variety of recovery mechanisms designed to correct the problem. The POWER6 processor-based systems include extensive hardware and firmware recovery logic.

Fault Isolation Registers

Error checker signals are captured and stored in hardware Fault Isolation Registers (FIRs). The associated *Who's on First* logic circuitry is used to limit the domain of an error to the first checker that encounters the error. In this way, run-time error diagnostics can be deterministic such that for every check station, the unique error domain for that checker is defined and documented. Ultimately, the error domain becomes the Field Replaceable Unit (FRU) call, and manual interpretation of the data is not normally required.

First Failure Data Capture (FFDC)

First Failure Data Capture (FFDC) is an error isolation technique that ensures that when a fault is detected in a system through error checkers or other types of detection methods, the root cause of the fault will be captured without the need to recreate the problem or run an extended tracing or diagnostics program.

For the vast majority of faults, a good FFDC design means that the root cause will be detected automatically without intervention of a service representative. Pertinent error data related to the fault is captured and saved for analysis. In hardware, FFDC data is collected from the fault isolation registers and 'Who's On First' logic. In Firmware, this data consists of return codes, function calls, etc.

FFDC “check stations” are carefully positioned within the server logic and data paths to ensure that potential errors can be quickly identified and accurately tracked to a Field Replaceable Unit (FRU).

This proactive diagnostic strategy is a significant improvement over the classic, less accurate “reboot and diagnose” service approaches.

Figure 4-8 shows a schematic of a Fault Isolation Register implementation.

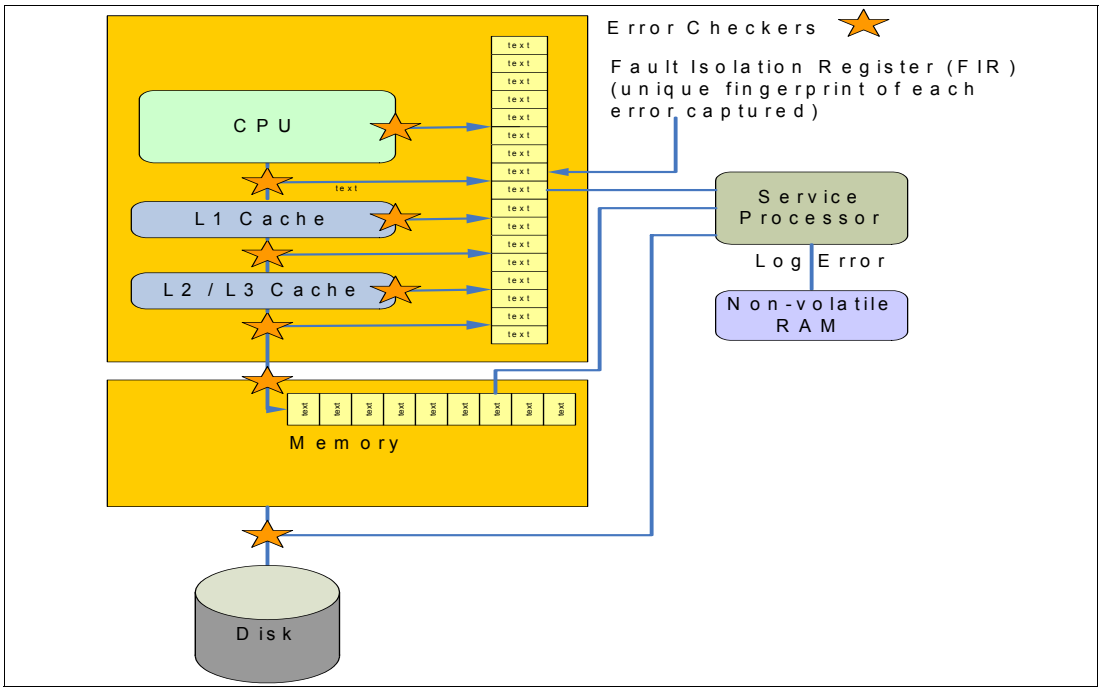


Figure 4-8 Schematic of a FIR implementation

Fault isolation

The Service Processor interprets error data captured by the FFDC checkers – saved in the FIRs and ‘Who’s On First logic or other firmware related data capture methods – in order to determine the root cause of the error event.

Root cause analysis may indicate that the event is recoverable, meaning that a service action point or need for repair has not been reached. Alternatively, it could indicate that a service action point has been reached, where the event exceeded a pre-determined threshold or was unrecoverable. Based upon the isolation analysis, recoverable error threshold counts may be incremented. No specific service action could be necessary when the event is recoverable.

When the event requires a service action, additional required information will be collected to service the fault. For unrecoverable errors or for recoverable events that meet or exceed their service threshold – meaning a service action point has been reached – a request for service will be initiated through an error logging component.

4.3.2 Diagnosing problems

Using the extensive network of advanced and complementary error detection logic built directly into hardware, firmware, and operating systems, the IBM POWER6 processor-based Systems can perform considerable self diagnosis.

Boot-time

When an IBM POWER6 processor-based system powers up, the Service Processor initializes system hardware. Boot-time diagnostic testing uses a multi-tier approach for system validation, starting with managed low-level diagnostics supplemented with system firmware initialization and configuration of I/O hardware, followed by OS-initiated software test routines. Boot-time diagnostic routines include:

- Built-in-Self-Tests (BISTs) for both logic components and arrays ensure the internal integrity of components. Because the Service Processor assist in performing these tests, the system is enabled to perform fault determination and isolation whether system processors are operational or not. Boot-time BISTs may also find faults undetectable by processor-based Power-on-Self-Test (POST) or diagnostics.
- Wire-Tests discover and precisely identify connection faults between components such as processors, memory, or I/O hub chips.
- Initialization of components such as ECC memory, typically by writing patterns of data and allowing the server to store valid ECC data for each location, can help isolate errors.

In order to minimize boot time, the system will determine which of the diagnostics are required to be started in order to ensure correct operation based on the way the system was powered off, or on the boot-time selection menu.

Runtime

All POWER6 processor-based systems can monitor critical system components during run-time, and they can take corrective actions when recoverable faults occur. IBMs hardware error check architecture provides the ability to report non-critical errors in an ‘out-of-band’ communications path to the Service Processor without affecting system performance.

A significant part of IBMs runtime diagnostic capabilities originate with the POWER6 Service Processor. Extensive diagnostic and fault analysis routines have been developed and improved over many generations of POWER processor-based servers, and enable quick and accurate predefined responses to both actual and potential system problems.

The Service Processor correlates and processes runtime error information, using logic derived from IBMs engineering expertise to count recoverable errors (called Thresholding) and predict when corrective actions must be automatically initiated by the system. These actions can include:

- Requests for a part to be replaced.
- Dynamic (on-line) invocation of built-in redundancy for automatic replacement of a failing part.
- Dynamic deallocation of failing components so that system availability is maintained.

Device drivers

In certain cases diagnostics are best performed by operating system-specific drivers, most notably I/O devices that are owned directly by a logical partition. In these cases, the operating system device driver will often work in conjunction with I/O device microcode to isolate and/or recover from problems. Potential problems are reported to an operating system device driver, which logs the error. I/O devices may also include specific exercisers that can be invoked by the diagnostic facilities for problem recreation if required by service procedures.

4.3.3 Reporting problems

In the unlikely event of a system hardware or environmentally induced failure is diagnosed, POWER6 processor-based systems report the error through a number of mechanisms. This ensures that appropriate entities are aware that the system may be operating in an error state. However, a crucial piece of a solid reporting strategy is ensuring that a single error communicated through multiple error paths is correctly aggregated, so that later notifications are not accidently duplicated.

Error logging and analysis

Once the root cause of an error has been identified by a fault isolation component, an error log entry is created with some basic data such as:

- An error code uniquely describing the error event
- The location of the failing component
- The part number of the component to be replaced, including pertinent data like engineering and manufacturing levels
- Return codes
- Resource identifiers
- First Failure Data Capture data

Data containing information on the effect that the repair will have on the system is also included. Error log routines in the operating system can then use this information and decide to call home to contact service and support, send a notification message, or continue without an alert.

Remote support

The Remote Management and Control (RMC) application is delivered as part of the base operating system, including the operating system running on the Hardware Management Console. RMC provides a secure transport mechanism across the LAN interface between the operating system and the Hardware Management Console and is used by the operating system diagnostic application for transmitting error information. It performs a number of other functions as well, but these are not used for the service infrastructure.

Manage serviceable events

A critical requirement in a logically partitioned environment is to ensure that errors are not lost before being reported for service, and that an error should only be reported once, regardless of how many logical partitions experience the potential effect of the error. The *Manage Serviceable Events* task on the Hardware Management Console (HMC) is responsible for aggregating duplicate error reports, and ensures that all errors are recorded for review and management.

When a local or globally reported service request is made to the operating system, the operating system diagnostic subsystem uses the Remote Management and Control Subsystem (RMC) to relay error information to the Hardware Management Console. For global events (platform unrecoverable errors, for example) the Service Processor will also forward error notification of these events to the Hardware Management Console, providing a redundant error-reporting path in case of errors in the RMC network.

The first occurrence of each failure type will be recorded in the *Manage Serviceable Events* task on the Hardware Management Console. This task will then filter and maintain a history of duplicate reports from other logical partitions or the Service Processor. It then looks across all active service event requests, analyzes the failure to ascertain the root cause and, if enabled,

initiates a call home for service. This methodology insures that all platform errors will be reported through at least one functional path, ultimately resulting in a single notification for a single problem.

Extended Error Data (EED)

Extended error data (EED) is additional data that is collected either automatically at the time of a failure or manually at a later time. The data collected is dependent on the invocation method but includes information like firmware levels, operating system levels, additional fault isolation register values, recoverable error threshold register values, system status, and any other pertinent data.

The data is formatted and prepared for transmission back to IBM to assist the service support organization with preparing a service action plan for the service representative or for additional analysis.

System dump handling

In some circumstances, an error may require a dump to be automatically or manually created. In this event, it will be offloaded to the HMC upon reboot. Specific HMC information is included as part of the information that can optionally be sent to IBM support for analysis. If additional information relating to the dump is required, or if it becomes necessary to view the dump remotely, the HMC dump record will notify IBMs support center upon which HMC the dump is located.

4.3.4 Notifying the appropriate contacts

Once a POWER6 processor-based system has detected, diagnosed, and reported an error to an appropriate aggregation point, it then takes steps to notify the customer, and if necessary the IBM Support Organization. Depending upon the assessed severity of the error and support agreement, this notification could range from a simple notification to having field service personnel automatically dispatched to the customer site with the correct replacement part.

Customer notify

When an event is important enough to report, but doesn't indicate the need for a repair action or the need to call home to IBM service and support, it is classified as *customer notify.* Customers are notified because these events might be of interest to an administrator. The event might be a symptom of an expected systemic change, such as a network reconfiguration or failover testing of redundant power or cooling systems. Examples of these events include:

- Network events like the loss of contact over a Local Area Network (LAN)
- Environmental events such as ambient temperature warnings
- Events that need further examination by the customer, but these events do not necessarily require a part replacement or repair action

Customer notify events are serviceable events by definition because they indicate that something has happened which requires customer awareness in the event they want to take further action. These events can always be reported back to IBM at the customer's discretion.

Call home

A correctly configured POWER6 processor-based system can initiate an automatic or manual call from a customer location to the IBM service and support organization with error data, server status, or other service-related information. Call home invokes the service organization

in order for the appropriate service action to begin, automatically opening a problem report and in some cases also dispatching field support. This automated reporting provides faster and potentially more accurate transmittal of error information. While configuring call home is optional, customers are strongly encouraged to configure this feature in order to obtain the full value of IBM service enhancements.

Vital Product Data (VPD) and inventory management

POWER6 processor-based systems store vital product data (VPD) internally, which keeps a record of how much memory is installed, how many processors are installed, manufacturing level of the parts, and so on. These records provide valuable information that can be used by remote support and service representatives, enabling them to provide assistance in keeping the firmware and software on the server up-to-date.

IBM problem management database

At the IBM support center, historical problem data is entered into the IBM Service and Support Problem Management database. All of the information related to the error along with any service actions taken by the service representative are recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

4.3.5 Locating and repairing the problem

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts requiring service. POWER6 processor-based systems utilize a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Guiding light LEDs

Guiding Light uses a series of flashing LEDs, allowing a service provider to quickly and easily identify the location of system components. Guiding Light can also handle multiple error conditions simultaneously which could be necessary in some very complex high-end configurations.

In the Guiding Light LED implementation, when a fault condition is detected on a POWER6 processor-based system, an amber System Attention LED will be illuminated. Upon arrival, the service provider engages *identify* mode by selecting a specific problem. The Guiding Light system then identifies the part that needs to be replaced by flashing the amber *identify* LED.

Datacenters can be complex places, and Guiding Light is designed to do more than identify visible components. When a component might be hidden from view, Guiding Light can flash a sequence of LEDs that extend to the frame exterior, clearly “guiding” the service representative to the correct rack, system, enclosure, drawer, and component.

The operator panel

The Operator Panel on a POWER6 processor-based system is a four row by 16 element LCD display used to present boot progress codes, indicating advancement through the system power-on and initialization processes. The Operator Panel is also used to display error and location codes when an error occurs that prevents the system from booting. The operator panel includes several buttons allowing a service representative or the customer to change various boot-time options, and perform a subset of the service functions that are available on the Advanced System Management Interface (ASMI).

Concurrent maintenance

The IBM POWER6 processor-based systems are designed with the understanding that certain components have higher intrinsic failure rates than others. The movement of fans, power supplies, and physical storage devices naturally make them more susceptible to wear down or burnout, while other devices such as I/O adapters may begin to wear from repeated plugging or unplugging. For this reason, these devices are specifically designed to be concurrently maintainable, when properly configured.

In other cases, a customer may be in the process of moving or redesigning a datacenter, or planning a major upgrade. At times like these, flexibility is crucial. The IBM POWER6 processor-based systems are designed for redundant or concurrently maintainable power, fans, physical storage, and I/O towers.

Blind-swap PCI adapters

Blind-swap PCI adapters represent significant service and ease-of-use enhancements in I/O subsystem design while maintaining high PCI adapter density.

Standard PCI designs supporting *hot-add* and *hot-replace* require top access so that adapters can be slid into the PCI I/O slots vertically. *Blind-swap* allows PCI adapters to be concurrently replaced without having to put the I/O drawer into a service position.

Firmware updates

Firmware updates for POWER6 processor-based systems are released in a cumulative sequential fix format, packaged as an RPM for concurrent application and activation. Administrators can install and activate many firmware patches without cycling power or rebooting the server.

The new firmware image is loaded on the HMC using any of the following methods:

- IBM-distributed media such as a CD-ROM
- A Problem Fix distribution from the IBM Service and Support repository
- Download from the IBM Web site:

`http://www14.software.ibm.com/webapp/set2/firmware/gjsn`

- FTP from another server

IBM will support multiple firmware releases in the field, so under expected circumstances a server can operate on an existing firmware release, using concurrent firmware fixes to stay up-to-date with the current patch level. Because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during system operation, a patch in this area will require a system reboot for activation. Under normal operating conditions, IBM intends to provide patches for an individual firmware release level for up to two years after first making the release code generally availability. After this period, clients should plan to update in order to stay on a supported firmware release.

Activation of new firmware functions, as opposed to patches, will require installation of a new firmware release Level. This process is disruptive to server operations in that it requires a scheduled outage and full server reboot.

In addition to concurrent and disruptive firmware updates, IBM will also offer concurrent patches that include functions which are not activated until a subsequent server reboot. A server with these patches will operate normally. The additional concurrent fixes will be installed and activated when the system reboots after the next scheduled outage.

Additional capability is being added to the POWER6 firmware to be able to view the status of a system power control network background firmware update. This subsystem will update as necessary as migrated nodes or I/O drawers are added to the configuration. The new firmware will not only provide an interface to be able to view the progress of the update, but also control starting and stopping of the background update if a more convenient time becomes available.

Repair and verify

Repair and Verify (R&V) is a system used to guide a service provider step-by-step through the process of repairing a system and verifying that the problem has been repaired. The steps are customized in the appropriate sequence for the particular repair for the specific system being repaired. Repair scenarios covered by repair and verify include:

- Replacing a defective Field Replaceable Unit (FRU)
- Reattaching a loose or disconnected component
- Correcting a configuration error
- Removing or replacing an incompatible FRU
- Updating firmware, device drivers, operating systems, middleware components, and IBM applications after replacing a part
- Installing a new part

Repair and verify procedures are designed to be used both by service representative providers who are familiar with the task at hand and those who are not. On Demand Education content is placed in the procedure at the appropriate locations. Throughout the repair and verify procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the Serviceable Event, to ensure that the guided maintenance procedures are operating correctly.

Service documentation on the support for IBM System p

The support for IBM System p Web site is an electronic information repository for POWER6 processor-based systems. This Web site also provides online training and educational material, as well as service documentation. In addition, the Web site will provide service procedures that are not handled by the automated Repair and Verify guided component.

The support for System p Web site is located at:

http://www.ibm.com/systems/support/p

Clients can subscribe through the Subscription Services to obtain the notifications on the latest updates available for service related documentation. The latest version of the documentation is accessible through the Internet, and a CD-ROM-based version is also available.

4.4 Operating System support for RAS features

Table 4-1 on page 114 gives an overview of a number of features for continuous availability supported by the different operating systems running on the POWER6 processor-based systems.

Table 4-1 Operating system support for selected RAS features

RAS feature	AIX V5.3	AIX V6.1	RHEL V5.1	SLES V10
System Deallocation of Failing Components				
Dynamic processor deallocation	Y	Y	Y[1]	Y
Dynamic processor sparing	Y	Y	Y	Y
Processor instruction retry	Y	Y	Y	Y
Alternate processor recovery	Y	Y	Y	Y
Partition contained checkstop	Y	Y	Y	Y
Persistent processor deallocation	Y	Y	Y	Y
GX+ bus persistent deallocation	Y	Y	N	N
PCI bus extended error detection	Y	Y	Y	Y
PCI bus extended error recovery	Y	Y	Limited[1]	Limited
PCI-PCI bridge extended error handling	Y	Y	N	N
Redundant RIO Link	Y	Y	Y	Y
PCI card hot swap	Y	Y	Y[1]	Y
Dynamic SP failover at runtime	Y	Y	N	N
Memory sparing with CoD at IPL time	Y	Y	Y	Y
Clock failover at IPL	Y	Y	Y	Y
Memory Availability				
ECC Memory, L2, L3 cache	Y	Y	Y	Y
Dynamic bit-steering (spare memory)	Y	Y	Y	Y
Memory scrubbing	Y	Y	Y	Y
Chipkill memory	Y	Y	Y	Y
Memory page deallocation	Y	Y	N	N
L1 parity check plus retry	Y	Y	Y	Y
L2 cache line delete	Y	Y	Y	Y
L3 cache line delete	Y	Y	Y	Y
L3 cache memory scrubbing	Y	Y	Y	Y
Array recovery & Array persistent deallocation (spare bits in L1 & L2 cache; L1, L2, and L3 directory)	Y	Y	Y	Y
Special uncorrectable error handling	Y	Y	Y	Y
Fault Detection and Isolation				
Platform FFDC diagnostics	Y	Y	Y	Y
I/O FFDC diagnostics	Y	Y	N	Y
Runtime diagnostics	Y	Y	Limited	Limited

RAS feature	AIX V5.3	AIX V6.1	RHEL V5.1	SLES V10
Storage protection keys	Y	Y	N	N
Dynamic trace	N	Y	N	N
Operating system FFDC	Y	Y	N	N
Error log analysis	Y	Y	Y	Y
Service processor support for BIST for logic & arrays, wire tests, and component initialization	Y	Y	Y	Y
Serviceability				
Boot time progress indicator	Y	Y	Limited	Limited
Firmware error codes	Y	Y	Y	Y
Operating system error codes	Y	Y	Limited	Limited
Inventory collection	Y	Y	Y	Y
Environmental and power warnings	Y	Y	Y	Y
Hot plug fans, power supplies	Y	Y	Y	Y
Extended error data collection	Y	Y	Y	Y
SP call home on non-HMC configurations	Y	Y	Y	Y
I/O drawer redundant connections	Y	Y	Y	Y
I/O drawer hot-add and concurrent repair	Y	Y	Y	Y
SP mutual surveillance with POWER Hypervisor	Y	Y	Y	Y
Dynamic firmware update with the HMC	Y	Y	Y	Y
Service agent call home application	Y	Y	Y	Y
Guiding light LEDs	Y	Y	Y	Y
System dump for memory, POWER Hypervisor, SP	Y	Y	Y	Y
Operating system error reporting to HMC SFP application	Y	Y	Y	Y
RMC secure error transmission subsystem	Y	Y	Y	Y
Health check scheduled operations with HMC	Y	Y	Y	Y
Operator panel (virtual or real)	Y	Y	Y	Y
Redundant HMCs	Y	Y	Y	Y
Automated recovery/restart	Y	Y	Y	Y
Repair and verify guided maintenance	Y	Y	Limited	Limited
Concurrent kernel update	N	Y	N	N

1 feature is not supported on Version 4 of RHEL

4.5 Manageability

Several functions and tools help manageability, and can allow you to efficiently and effectively manage your system.

4.5.1 Service processor

The service processor is a controller running its own operating system. It is a component of the service interface card.

The service processor operating system has specific programs and device drivers for the service processor hardware. The host interface is a processor support interface connected to the POWER6 processor. The service processor is always working, regardless of main system unit's state. The system unit can be in the following states:

- Standby (Power off)
- Operating, ready to start partitions
- Operating with running logical partitions

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, ensuring the connection to the HMC for manageability purposes and accepting Advanced System Management Interface (ASMI) Secure Sockets Layer (SSL) network connections. The service processor provides the ability to view and manage the machine-wide settings using the ASMI, and allows complete system and partition management from the HMC.

Note: The service processor enables a system that will not boot to be analyzed. The error log analysis can be performed from either the ASMI or the HMC.

The service processor uses two Ethernet 10/100 Mbps ports:

- Both Ethernet ports are only visible to the service processor and can be used to attach the server to an HMC or to access the ASMI. The ASMI options can be accessed through an HTTP server that is integrated into the service processor operating environment.
- Both Ethernet ports have a default IP address:
  - Service processor Eth0 or HMC1 port is configured as 169.254.2.147 (This applies to the service processor in drawer 1 or the top drawer.)
  - Service processor Eth1 or HMC2 port is configured as 169.254.3.147 (This applies to the service processor in drawer 1 or the top drawer.)
  - Service processor Eth0 or HMC1 port is configured as 169.254.2.146 (This applies to the service processor in drawer 2 or the second drawer from top to bottom.)
  - Service processor Eth0 or HMC1 port is configured as 169.254.3.146 (This applies to the service processor in drawer 2 or the second drawer from top to bottom.)

4.5.2 System diagnostics

The system diagnostics consist of stand-alone diagnostics, which are loaded from the DVD-ROM drive, and online diagnostics (available in AIX).

- Online diagnostics, when installed, are a part of the AIX operating system on the disk or server. They can be booted in single-user mode (service mode), run in maintenance

mode, or run concurrently (concurrent mode) with other applications. They have access to the AIX error log and the AIX configuration data.

- Service mode, which requires a service mode boot of the system, enables the checking of system devices and features. Service mode provides the most complete checkout of the system resources. All system resources, except the SCSI adapter and the disk drives used for paging, can be tested.
- Concurrent mode enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, some devices might require additional actions by the user or diagnostic application before testing can be done.
- Maintenance mode enables the checking of most system resources. Maintenance mode provides the same test coverage as service mode. The difference between the two modes is the way they are invoked. Maintenance mode requires that all activity on the operating system be stopped. The **shutdown -m** command is used to stop all activity on the operating system and put the operating system into maintenance mode.

- The System Management Services (SMS) error log is accessible on the SMS menus. This error log contains errors that are found by partition firmware when the system or partition is booting.
- The service processor's error log can be accessed on the ASMI menus.
- You can also access the system diagnostics from a Network Installation Management (NIM) server.

Note: Because the 570 system have an optional DVD-ROM (FC 5756) and DVD-RAM (FC 5757), alternate methods for maintaining and servicing the system need to be available if you do not order the DVD-ROM or DVD-RAM

4.5.3 Electronic Service Agent

Electronic Service Agent™ and the IBM Electronic Services Web portal comprise the IBM Electronic Service solution. IBM Electronic Service Agent is a no-charge tool that proactively monitors and reports hardware events, such as system errors, performance issues, and inventory. Electronic Service Agent can help focus on the customer's company strategic business initiatives, save time, and spend less effort managing day-to-day IT maintenance issues.

Now integrated in AIX 5L V5.3 TL6 in addition to the HMC, Electronic Service Agent is designed to automatically and electronically report system failures and customer-perceived issues to IBM, which can result in faster problem resolution and increased availability. System configuration and inventory information collected by Electronic Service Agent also can be viewed on the secure Electronic Service web portal, and used to improve problem determination and resolution between the customer and the IBM support team. As part of an increased focus to provide even better service to IBM customers, Electronic Service Agent tool configuration and activation comes standard with the system. In support of this effort, a new HMC External Connectivity security whitepaper has been published, which describes data exchanges between the HMC and the IBM Service Delivery Center (SDC) and the methods and protocols for this exchange.

To access Electronic Service Agent user guides, perform the following steps:

1. Go to the IBM Electronic Services news Web site at

   https://www-304.ibm.com/jct03004c/support/electronic/portal

2. Select your country
3. Click “IBM Electronic Service Agent Connectivity Guide”

Note: To receive maximum coverage, activate Electronic Service Agent on every platform, partition, and Hardware Management Console (HMC) in your network. If your IBM System p server is managed by an HMC, the HMC will report all hardware problems, and the AIX operating system will report only software problems and system information. You must configure the Electronic Service Agent on the HMC. The AIX operating system will not report hardware problems for a system managed by an HMC.

IBM Electronic Service provide these benefits:

Increased uptime — Electronic Service Agent is designed to enhance the warranty and maintenance service by providing faster hardware error reporting and uploading system information to IBM support. This can optimize the time monitoring the symptoms, diagnosing the error, and manually calling IBM support to open a problem record. 24x7 monitoring and reporting means no more dependency on human intervention or off-hours customer personnel when errors are encountered in the middle of the night.

Security — Electronic Service Agent is secure in monitoring, reporting, and storing the data at IBM. Electronic Service Agent securely transmits via the internet (HTTPS or VPN) and can be configured to communicate securely through gateways to provide customers a single point of exit from their site. Communication between the customer and IBM only flows one way. Activating Service Agent does not enable IBM to call into a customer’s system. System inventory information is stored in a secure database, which is protected behind IBM firewalls. The customer’s business applications or business data is never transmitted to IBM

More accurate reporting — Since system information and error logs are automatically uploaded to the IBM support Center in conjunction with the service request, customers are not required to find and send system information, decreasing the risk of misreported or misdiagnosed errors. Once inside IBM, problem error data is run through a data knowledge management system and knowledge articles are appended to the problem record.

Customized support — Using the IBM ID entered during activation, customers can view system and support information in the “My Systems” and “Premium Search” sections of the Electronic Services Web site.

The Electronic Services Web portal is a single Internet entry point that replaces the multiple entry points traditionally used to access IBM Internet services and support. This Web portal enables you to gain easier access to IBM resources for assistance in resolving technical problems.

Service Agent provides these additional services:

- My Systems: Client and IBM employees authorized by the client can view hardware and software information and error messages that are gathered by Service Agent on Electronic Services Web pages at:

  https://www-304.ibm.com/jct03004c/support/electronic/portal

Premium Search: A search service using information gathered by Service Agents (this is a paid service that requires a special contract).

For more information on how to utilize the power of IBM Electronic Services, visit the following Web site or contact an IBM Systems Services Representative.

https://www-304.ibm.com/jct03004c/support/electronic/portal

4.5.4 Manage serviceable events with the HMC

Service strategies become more complicated in a partitioned environment. The Manage Serviceable Events task in the HMC can help streamline this process.

Each logical partition reports errors it detects, without determining whether other logical partitions also detect and report the errors. For example, if one logical partition reports an error for a shared resource, such as a managed system power supply, other active logical partitions might report the same error.

By using the Manage Serviceable Events task in the HMC, you can avoid long lists of repetitive call-home information by recognizing that these are repeated errors and consolidating them into one error.

In addition, you can use the Manage Serviceable Events task to initiate service functions on systems and logical partitions including the exchanging of parts, configuring connectivity, and managing dumps.

4.5.5 Hardware user interfaces

In addition, you can use the Manage Serviceable Events task to initiate service functions on systems and logical partitions including the exchanging of parts, configuring connectivity, and managing dumps.

Advanced system Management Interface

The Advanced System Management interface (ASMI) is the interface to the service processor that enables you to manage the operation of the server, such as auto power restart, and to view information about the server, such as the error log and vital product data. Some repair procedures require connection to the ASMI.

The ASMI is accessible through the HMC. For details, see “Accessing the ASMI using an HMC” The ASMI is also accessible using a Web browser on a system that is connected directly to the service processor (in this case, either a standard Ethernet cable or a crossed cable) or through an Ethernet network. Use the ASMI to change the service processor IP addresses or to apply some security policies and avoid the access from undesired IP addresses or range.

You might be able to use the service processor’s default settings. In that case, accessing the ASMI is not necessary

Accessing the ASMI using an HMC

If configured to do so, the HMC connects directly to the ASMI for a selected system from this task.

To connect to the Advanced System Management interface from an HMC :

1. Open Systems Management from the navigation pane.
2. From the work pane, select one or more managed systems to work with.

3. From the System Management tasks list, select Operations.
4. From the Operations task list, select Advanced System Management Interface (ASMI).

Accessing the ASMI using a Web browser

The Web interface to the ASMI is accessible through Microsoft Internet Explorer 6.0, Microsoft Internet Explorer 7, Netscape 7.1, Mozilla Firefox, or Opera 7.23 running on a PC or mobile computer connected to the service processor. The Web interface is available during all phases of system operation, including the initial program load (IPL) and run time. However, some of the menu options in the Web interface are unavailable during IPL or run time to prevent usage or ownership conflicts if the system resources are in use during that phase. The ASMI provides a Secure Sockets Layer (SSL) Web connection to the service processor. To establish an SSL connection, open your browser using “https://”.

Note: To make the connection thorough Internet Explorer, click “Tools”, “Internet Options”. Uncheck “Use TLS 1.0”, and click OK.

Accessing the ASMI using an ASCII terminal

The ASMI on an ASCII terminal supports a subset of the functions provided by the Web interface and is available only when the system is in the platform standby state. The ASMI on an ASCII console is not available during some phases of system operation, such as the IPL and run time

Graphics terminal

The graphics terminal is available to users who want a graphical user interface (GUI) to their AIX or Linux systems. To use the graphics terminal, plug the graphics adapter into a PCI slot in the back of the server. You can connect a standard monitor, keyboard, and mouse to the adapter to use the terminal. This connection allows you to access the SMS menus, as well as an operating system console.

4.5.6 IBM System p firmware maintenance

The IBM System p, IBM System p5, pSeries, and RS/6000 Client-Managed Microcode is a methodology that enables you to manage and install microcode updates on IBM System p, IBM System p5, pSeries, and RS/6000 systems and associated I/O adapters. The IBM System p microcode can be installed either from an HMC or from a running partition in case that system is not managed by an HMC. For update details, see below web page.

http://www14.software.ibm.com/webapp/set2/firmware/gjsn

If you use an HMC to manage your server, you can use the HMC interface to view the levels of server firmware and power subsystem firmware that are installed on your server and are available to download and install.

Each IBM System p server has the following levels of server firmware and power subsystem firmware:

- Installed level – This is the level of server firmware or power subsystem firmware that has been installed and will be installed into memory after the managed system is powered off and powered on. It is installed on the *t* side of system firmware.
- Activated level – This is the level of server firmware or power subsystem firmware that is active and running in memory.

- Accepted level – This is the backup level of server or power subsystem firmware. You can return to this level of server or power subsystem firmware if you decide to remove the installed level. It is installed on the *p* side of system firmware.

IBM provide the Concurrent Firmware Maintenance (CFM) function on System p systems. This function supports applying nondisruptive system firmware service packs to the system concurrently (without requiring a reboot to activate changes). For systems that are not managed by an HMC, the installation of system firmware is always disruptive.

The concurrent levels of system firmware can, on occasion, contain fixes that are known as *deferred*. These deferred fixes can be installed concurrently but are not activated until the next IPL. For deferred fixes within a service pack, only the fixes in the service pack, which cannot be concurrently activated, are deferred. Figure 4-9 shows the system firmware file naming convention.

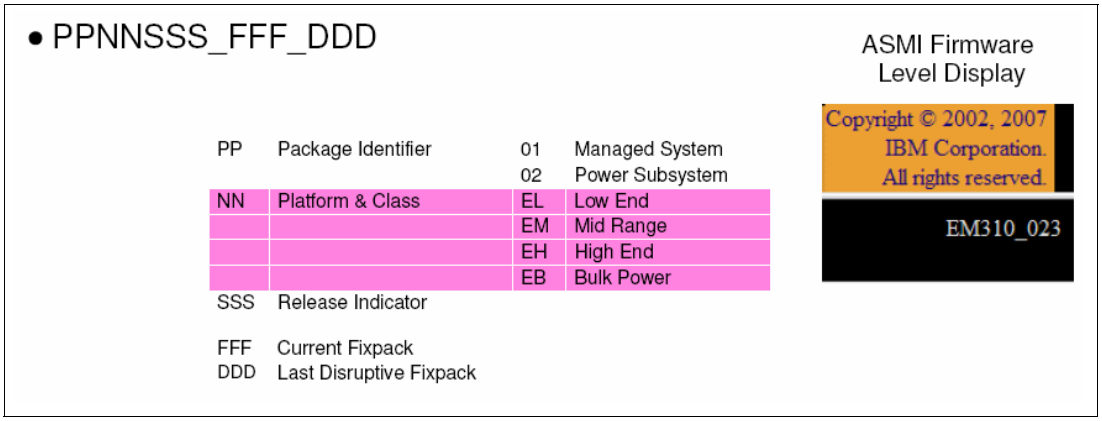


Figure 4-9 firmware file naming convention

Here is one example.

Example 4-1

```
01EM310_026_026 = Managed System Firmware for 9117-MMA Release 310 Fixpack 026
```

An installation is disruptive if:

- The release levels (*SSS*) of currently installed and new firmware are different.
- The service pack level (*FFF*) and the last disruptive service pack level (*DDD*) are equal in new firmware.

Otherwise, an installation is concurrent if:

- The service pack level (*FFF*) of the new firmware is higher than the service pack level currently installed on the system and the above conditions for disruptive installation are not met.

4.5.7 Management Edition for AIX

IBM Management Edition for AIX (ME for AIX) is designed to provide robust monitoring and quick time to value by incorporating out-of-the box best practice solutions that were created by AIX and PowerVM Virtual I/O Server developers. These best practice solutions include predefined thresholds for alerting on key metrics, Expert Advice that provides an explanation of the alert and recommends potential actions to take to resolve the issue, and the ability to

take resolution actions directly from the Tivoli Enterprise Portal or set up automated actions. Users have the ability to visualize the monitoring data in the Tivoli Enterprise Portal determine the current state of the AIX, LPAR, CEC, HMC and VIOS resources.

Management Edition for AIX is an integrated systems management offering created specifically for the system p platform that provides as primary functions :

- Monitoring of the health and availability of the system p.
- Discovery of configurations and relationships between system p service and application components.
- Usage and accounting of system p IT resources.

For information regarding the ME for AIX, visit the following link :

http://www-03.ibm.com/systems/p/os/aix/sysmgmt/me/index.html

4.5.8 IBM Director

IBM Director is an integrated, easy-to-use suite of tools that provide you with flexible system management capabilities to help realize maximum systems availability and lower IT costs.

IBM Director provide below benefits :

- An easy-to-use, integrated suite of tools with consistent look-and-feel and single point of management simplifies IT tasks
- Automated, proactive capabilities that help reduce IT costs and maximize system availability
- Streamlined, intuitive user interface to get started faster and accomplish more in a shorter period of time
- Open, standards-based design and broad platform and operating support enable customers to manage heterogeneous environments from a central point
- Can be extended to provide more choice of tools from the same user interface

For information regarding the IBM Director, visit the following links :

http://www-03.ibm.com/systems/management/director/

4.6 Cluster solution

Today's IT infrastructure requires that servers meet increasing demands, while offering the flexibility and manageability to rapidly develop and deploy new services. IBM clustering hardware and software provide the building blocks, with availability, scalability, security, and single-point-of-management control, to satisfy these needs. The advantages of clusters are:

- High processing capacity
- Resource consolidation
- Optimal use of resources
- Geographic server consolidation
- 24x7 availability with failover protection
- Disaster recovery
- Scale-out and scale-up without downtime

- Centralized system management

The POWER processor-based AIX and Linux cluster target scientific and technical computing, large-scale databases, and workload consolidation. IBM Cluster Systems Management software (CSM) is designed to help reduce the overall cost and complexity of IT management by simplifying the tasks of installing, configuring, operating, and maintaining clusters of servers or logical partitions (LPARs). CSM offers a single consistent interface for managing both AIX and Linux nodes, with capabilities for remote parallel network installation, remote hardware control, distributed command execution, file collection and distribution, cluster-wide monitoring capabilities, and integration with High Performance Computing (HPC) applications.

CSM V1.7 which is need to support POWER6 processor-based HMC include Highly Available Management Server (HA MS) at no additional charge. CSM HAMS is positioned for enterprises that need a highly available management server. CSM HA MS is designed to remove the management server as a single point of failure in the cluster.

For information regarding the IBM Cluster Systems Management for AIX , HMC control, cluster building block servers, and cluster software available, visit the following links:

- Cluster 1600

  http://www-03.ibm.com/systems/clusters/hardware/1600/index.html

- Cluster 1350™

  http://www-03.ibm.com/systems/clusters/hardware/1350/index.html

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this Redpaper.

IBM Redbooks

For information about ordering these publications, see “How to get Redbooks” on page 126. Note that some of the documents referenced here may be available in softcopy only.

- *PowerVM Virtualization on IBM System p Introduction and Configuration Forth Edition*, SG24-7940
- *PowerVM Virtualization on IBM System p Managing and Monitoring*, SG24-7590
- *Getting started with PowerVM Lx86*, REDP-4298
- *IBM System p Live Partition Mobility*, SG24-7460
- *Integrated Virtualization Manager on IBM System p5*, REDP-4061
- *Introduction to Workload Partition Management in IBM AIX Version 6.1*, SG24-7431
- *Hardware Management Console V7 Handbook*, SG24-7491
- *LPAR Simplification Tools Handbook*, SG24-7231
- *IBM System p520 Technical Overview and Introduction*, REDP-4403
- *IBM System p550 Technical Overview and Introduction*, REDP-4404

Other publications

These publications are also relevant as further information sources for planning:

- *Logical Partitioning Guide*, SA76-0098
- *Site and Hardware Planning Guide*, SA76-0091
- *Site Preparation and Physical Planning Guide*, SA76-0103

These publications are also relevant as further information sources for installing:

- *Installation and Configuration Guide for the HMC*, SA76-0084
- *PCI Adapter Placement*, SA76-0090

These publications are also relevant as further information sources for using your system:

- *Introduction to Virtualization*, SA76-0145
- *Operations Guide for the ASMI and for Nonpartitioned Systems*, SA76-0094
- *Operations Guide for the HMC and Managed Systems*, SA76-0085
- *Virtual I/O Server Command Reference*, SA76-0101

These publications are also relevant as further information sources for troubleshooting:

- *AIX Diagnostics and Service Aids*, SA76-0106

© Copyright IBM Corp. 2008. All rights reserved.

- *Managing Devices*, SA76-0107
- *Managing PCI Devices*, SA76-0092
- *SAS RAID Controller Reference Guide*, SA76-0112
- *Service Guide for HMC Models 7042-Cr4 and 7042-C06*, SA76-0120

Online resources

These Web sites are also relevant as further information sources:

- IBM Systems Information Center

  http://publib.boulder.ibm.com/infocenter/systems
- Support for IBM System p

  http://www.ibm.com/systems/support/p
- IBM System Planning Tool

  http://www.ibm.com/systems/support/tools/systemplanningtool
- Fix Central / AIX operating system maintenance packages downloads

  http://www.ibm.com/eserver/support/fixes
- Microcode downloads

  http://www14.software.ibm.com/webapp/set2/firmware/gjsn
- Linux for IBM System p

  http://www.ibm.com/systems/p/linux/
- News on new computer technologies

  http://www.ibm.com/chips/micronews

How to get Redbooks

You can search for, view, or download Redbooks, Redpapers, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM Power 570 Technical Overview and Introduction

Expandable modular design supporting advanced mainframe class continuous availability enhancements

PowerVM virtualization including the optional Enterprise edition

POWER6 processor efficiency operating at state-of-the-art throughput levels

This IBM® Redpaper is a comprehensive guide covering the IBM System p 570 UNIX® server. The goal of this paper is to open the doors to the innovative IBM System p 570. It introduces major hardware offerings and discusses their prominent functions.

- Unique modular server packaging
- The new POWER6 processor available at frequencies of 3.5 GHz, 4.2 GHz, and 4.7 GHz.
- The specialized POWER6 DDR2 memory that provides greater bandwidth, capacity, and reliability.
- The new 1 Gb or 10 Gb Integrated Virtual Ethernet adapter that brings native hardware virtualization to this server
- PowerVM Live Partition Mobility
- Redundant service processor s to achieve continuous availability

This Redpaper expands the current set of IBM System p™ documentation by providing a desktop reference that offers a detailed technical description of the 570 system.

This Redpaper does not replace the latest marketing materials and tools. It is intended as an additional source of information that, together with existing materials, may be used to enhance your knowledge of IBM server solutions.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

REDP-4405-00